Letting the Public In

Opportunities and Standards for Open Data on
Beneficial Ownership, Country-by-Country Reporting
and Automatic Exchange of Financial Information

**FINANCIAL
TRANSPARENCY
COALITION**

AUTHORS

Tim Davies, Robin Gower

**Corresponding author:
tim.davies@opendataservices.coop**

ABOUT THE AUTHORS

Open Data Services is a UK-based workers cooperative, founded in 2015 to support organisations across the world to publish and use open data effectively.

Tim Davies is a co-founder of Open Data Services Cooperative, and a researcher focussed on the social impact of open data. He holds an MSc in Social Science of the Internet from the Oxford Institute, and is an affiliate and former fellow of the Harvard Berkman Center for Internet and Society. He is coauthor of the Open Contracting Data Standard, and from 20132015 was research coordinator for the World Wide Web Foundation's research network 'Exploring the Emerging Impacts of Open Data in Developing Countries'.

Robin Gower is an economist and software developer 10 years experience providing advice and guidance on data. He is a founding member of Open Data Manchester and is retained by the Big Lottery Fund as an economic, statistical and data expert.

Letting the Public In

Opportunities and Standards for Open Data on Beneficial Ownership, Country-by-Country Reporting and Automatic Exchange of Financial Information

---

Executive Summary

The Financial Transparency Coalition transparency platform outlines three core issues in need of greater transparency to curb the abuse of the 'shadow financial system': Beneficial Ownership, Country-by-Country Reporting and Automatic Exchange of Financial Information

Open data can be a powerful tool to secure effective transparency in each of these areas. A wide range of users, from governments and civil society activists, to private sector and individuals, have an interest in the reliable supply of high quality open data on the ultimate ownership of companies, where profits come from, and where the resulting wealth goes. In each of these areas, new data infrastructures are being developed-establishing new reporting regimes for information and new practices for the management of data. Standards play an important role in determining who the resulting data will be accessible to, whose needs it will meet, and how the costs of publishing and using the data will be distributed.

Towards gold standards?

There are many factors involved in providing a 'gold standard' for structured, accessible open data. These include: global interoperability; provision of unique identifiers; re-use of common data types; use of common encoding; approaches supporting internationalisation; provision of accessible documentation; mechanisms for feasible data validation; verifiability of published data; immutability of published data; balancing expressiveness and simplicity; open licensing; and robust governance .

None of the existing standardisation efforts with respect to Beneficial Ownership, Country-by-Country Reporting and Automatic Exchange of Financial Information provides a comprehensive gold standard for open data publication, nor is there a clear short-term path to the emergence of stand-alone gold standards. Instead, an approach is needed that draws upon common building blocks to help construct an ecosystem of standards and data. This must pay attention to standards being put forward for data collection and transfer, as these will substantially affect the nature of data that is subsequently available for open publication.

Costs of adoption

The costs of data standards adoption rest on a number of factors, including institutional frameworks, existing reporting practice, software and tooling, and the scale and scope of data management envisaged. In general, developing countries have less mature digital systems for managing key data, and so will face greater costs to establish the core systems on which open data publication can be based.

However, the current use of advanced data standards is not restricted to developed countries.

It is particularly notable that use of the XBRL standard for company reporting is in use in eleven of the top twenty secrecy jurisdictions.

Ultimately, standards affect the distribution of costs between data publishers and users. It is important to factor in the costs of not standardising, and to consider who benefits most from standards when looking at funding models to secure a robust open data infrastructure.

Strategic opportunities

Making standardisation work will require ongoing engagement at the policy and technology level, adopting a range of strategies to promote user focused open data, including:

1. **Engaging with open government and open data policy processes** to secure commitments to publish key data, particularly company registers and BO data;

2. **Critically engaging with existing standardisation processes** around CbCR and AEOI, paying close attention to the technical and policy decisions made, the governance of standards, and the engagement of users in testing the accessibility of the models proposed;

3. **Supporting the development of lightweight and prototype standards and data aggregation platforms** particularly for BO and direct publication of CbCR by companies. Such interventions can help shape the debate, providing a demonstration of the value of collecting specific data fields, adopting particular data structures, or pushing for an open-by-default policy with respect to key data.

Through the Open Government Partnership and International Open Data Charter process, there are a number of openings to pursue both political commitments to increased data publication that can support the FTC transparency platform, and collaboration on technical standardisation.

Equipped for advocacy

Although not identifying specific 'gold standards' this report seeks to provide a broad survey of the open data and standards landscape, going in-depth on key standards that are likely to surface in discussions of BO, CbCR and AEOI, and providing critical analysis of these. It offers a set of conceptual tools and key questions to consider in future advocacy engagement, flagging up specific points where the technical details really matter.

The transparency platform areas each have their own dynamics, and different strategies, alliances and interventions will be required to advance the state of the art in each.

Contents

Introduction

The Financial Transparency Coalition (FTC) is a global network of civil society, governments, and experts working to curtail illicit financial flows through the promotion of a transparent, accountable, and sustainable financial system that works for everyone[1]. FTC operates on the basis of a transparency platform, setting out three core issues in need of greater effective transparency in order to curb the abuse of the 'shadow financial system'. These three issues, Beneficial Ownership (BO), Country-by-Country Reporting (CbCR) and Automatic Exchange of Financial Information (AEOI) are underpinned by work on the international institutional architecture, enablers of illicit financial flows, and, more recently, a focus on the role of open data.

This report intends to explore role of open data standards. The brief for this report calls for:

- Identification of **the range of open data standards, which** could accompany the FTC transparency platform, including 'gold' standard options;
- Identification of related data **standards already in use by governments**, or to be used, in collecting data in FTC transparency platform areas, including comparing those used in developed and developing countries;
- **Recommendations on appropriate open data standards for** the transparency platform that are consistent with technological advances and which are able to be practically implemented worldwide;
- Assessment of the **political opportunities and challenges** to progressing the identified open data standards;
- Estimates of **the cost of countries transitioning** to systems that can produce relevant standardised (open) data;

Report Overview

To address these objectives we first turn to explore the transparency platform in more detail, identifying some of the critical issues that any data standards will need to address. This is followed by a set of vignettes outlining a series of use cases for data in each of these areas. We then unpack the different elements involved in data standardisation, drawing out key considerations for 'gold standard' options. The next section details existing standards, both open data standards, and the standards in use by governments and other institutions. This provides the basis for recommendations on standards, and a discussion of political opportunities and challenges. Cost estimates are summarised, with more detail provided in the Appendix.

Methodology

This study has involved consultation with partners across the Financial Transparency Coalition network, key informant interviews, a desk-based review of existing standards, and a workshop at the 6th Financial Transparency Conference in Jakarta. These responses and findings were then the subject of a political analysis and assessment of strategic options available to the Coalition.

....

1 https://financialtransparency.org/about/

Open data & standards

Over the last five years, interest in open data as a tool in transparency and accountability policy has grown. Open data standards have been developed in many different domains, and take many different forms.

What is open data?

The Open Knowledge Foundation's "Open Definition"[2] states that

> *"Open means anyone can freely access, use, modify, and share [data]*
> *for any purpose (subject, at most, to requirements that preserve*
> *provenance and openness)".*

This supports the wide distribution and reuse of data, free from technical and legal impediment. The open data movement, emerging in the late 2000s, has advocated for the proactive publication of substantial quantities of government held data, on everything from budgets, to transit schedules, to pollution statistics. A driving idea behind this is that, when users do not need to seek advance permission to use data, and can access it at marginal cost (usually zero for digital data), then usage increases, and many opportunities for innovative reuse of data are unlocked.

What are standards?

Standards describe how data should be prepared and published. They set out the technical features of a dataset, and the semantics of key fields and codes, making it possible to combine datasets from different sources and to build tools that know how to deal with predictable data. Standards can also set out criteria for how data should be made available to remove practical and legal barriers to data reuse. Standards are created by a wide range of different bodies, and their governance affects how often they will be changed or updated.

Individual standards do not stand alone. There are multiple layers of standardisation: and standards are almost always dependent upon other standards - creating an 'information infrastructure' which becomes more embedded and taken-for-granted over time.

In assessing open data standards for the FTC transparency platform, it is important to consider:

- **Definitions, vocabularies and taxonomies:** many key concepts within of the FTC transparency platform are defined in law, or given meaning through internationally agreed accounting standards. These key concepts may be defined in documents, but data standards will rely upon structured machine-readable taxonomies that can provide codes, definitional notes, and translations of these key concepts.

- **Technical formats and schema:** a wide range of file formats and structures are available for data representation, from CSV for tabular data, to XML for tree-structured data (see box-out below). The choice of format affects how clearly certain kinds of relationships can be expressed, and how easy the data will be to use in existing tools. For example, tabular data is easy to read in spreadsheet software, but can not represent complex relationships (such as ownership hierarchies) intuitively. XML

---

2 Open Definition: http://opendefinition.org/

supports more complex data structures, but requires specific expertise to work with. A schema define the specific fields that a file should use, setting out details of data representation and supporting the validation of data. The schema plays a central role in making data interoperable.

- **Licensing of standards and data:** data formats, schema and identifiers can all be openly licensed or proprietary, as can the data published using data standards. To qualify as an open standard, or as open data, standards and data should be provided under legal terms permitting free re-use- usually through the application of an explicit license.

- **Governance:** most standards change over time-adapting to incorporate new features, responding to publisher and user needs. The governance structure of a standard affects how changes can be made- and whether the standard can be unilaterally modified by a single body, or whether changes must be approved through a broad consultative process. The Open Stand Principles[3] set out a paradigm for open standards development.

- **Tooling and community:** software tools are important to support use and adoption of a data standard. If there are no tools that work with a standard, it is harder for publishers to know when their data is good quality, and for users to work with the data. Existing tools and an active user community, are important aspects of standard adoption.

Open data standards are often thought about as a way of structuring the publication of 'raw data'. Yet, in practice, they also end up structuring the collection of data inside institutions, as systems are adapted to produce the data requested by a standard, and data structures of an open standard are adopted. This makes advocacy for transparency standards a matter not only of what should be published, but also a matter of how data should be managed.

Standards in practice

The following examples illustrate a range of different approaches to securing adoption of a standard.

- **GTFS:** The General Transit Feed Specification (GTFS) is the defacto schema for describing transport timetables (in CSV). Created by the City of Portland, in collaboration with Google, it is widely used by public and private sector transport providers to get their services represented on platforms such as Google Maps. There is also a broad ecosystem of open source tools[4] (such as Open Trip Planner5[5]) based upon the data standard. Together these applications create a strong incentive for publishers to create data that complies with the standard indeed over 950 transit agencies worldwide already do so.

- **IATI:** The International Aid Transparency Initiative (IATI) standard provides an XML schema and a set of codelists (taxonomies) for publishing information on aid activities. The codelists used in the standard build heavily on documents published by the OECD Development Assistance Committee, but translate these into machine-readable vocabularies. An annual Aid Transparency Index, which ranks the quality of data,

---

3 https://open-stand.org/about-us/principles/
4 GTFS Tools: https://code.google.com/archive/p/googletransitdatafeed/wikis/OtherGTFSTools.wiki
5 Open Trip Planner: http://www.opentripplanner.org/

published by donors against the standard, and a technical secretariat charged with supporting publishers, have both been important elements in driving adoption. The technical secretariat also manages a regular upgrade process, making revisions to the standard. There has been some criticism of the complexity of the XML schema from NGOs, although tools have been created to hide some of the complexity of publishing.

- **AgroVOC:** The Agricultural Vocabulary (AgroVOC) is a long-established multilingual taxonomy maintained by the Food and Agriculture Organisation of the UN (FAO). An early adopter of Linked Data, the platform provides codes and URIs that can be used to refer to over 32,000 agricultural concepts. Crosswalks exist between AgroVOC and other widely used taxonomies. The FAO have invested in supporting the emergence of tools and plugins for existing software encourage uptake of AgroVOC standards.

- **EITI:** The Extractive Industry Transparency Initiative (EITI) data standards describe the kinds of things that EITI member countries should disclose but stop short of setting out a single schema for the machine-readable representation of this data. This focuses attention of definitions and processes, rather than schema and formats. Recently, EITI have started providing spreadsheet templates for data, although these lack clear data modeling and validation features commonly found in an open data standard.

- **OCDS:** The Open Contracting Data Standard (OCDS) provides a JSON schema and a suggested approach to publish data about all stages of a contracting process, from planning and tender, through to contract award and execution. It recommends publication of a permanent 'release' of structured open data for every change over the lifetime of the contracting process: offering the possibility of a robust record of change over time. Launched in 2014, tooling for working with OCDS data is only just starting to develop. Experience from early adopters of the standard shows that it is encouraging governments to look at linking-up between different systems that manage data on different stages of the procurement process.

Standards and structure: worked example

What difference do standards make? This worked example illustrates how standards can be applied to add structure to data.

The Tax Justice Network have called for authorities engaged in Automatic Exchange of Information to publish aggregate statistics from their data. The information requested includes: the number of accounts, by type, held by citizens from each jurisdiction; the number of accounts opened and closed in a reporting period; and the jurisdictions where intermediate companies (NFFE[6]) are resident.

In a world without data standards, each authority might report this information in their own way, in PDF reports, lists and tables. However, TJN have put forward a spreadsheet template for this information, an example of which is given below.

Jurisdiction of Residence (of Account Holder/CP)	Type of FI	Accounts of Controlling Persons (CPs) of Passive NFEs									
		Aggregate Value	Media Account Value	Aggregate Income	Median Account Income	# of CPs	# of Accounts	# of Accounts closed	Accounts [...] between 2013 & 2015		List of Jurisdictions where any Passive NFE is resident
									Opened	Closed	
ARGENTINA	Depository Institutions					1	1				Cayman Isl.

This starts moving towards structured data but it doesn't provide a robust template for the exchange of machine-readable data. First steps towards standardization of this as tabular data might involve:

- Replacing words with codes, each of which can have a label in multiple languages;
- Flattening[7] the data so that there is column for each 'dimension' of the data, and only one significant measure for each row;
- Making all implicit dimensions explicit;
- Specifying validation rules for the dimensions and measures;

For example, the first and last few columns of the table above would become:

Jurisdiction	Reporting Date (Start)	Reporting Data (End)	Currency	Measure	Value
AR	20130302	20150301	USD	aggregate_value	10000000
AR	20130302	20150301		accounts_openned	10
AR	20130302	20150301		accounts_closed	5
AR	20130302	20150301		jurisdictions_of_NFEEs	KY, BR

....

6 NonFinancial Foreign Entity

7 For more information on this 'normalisation' see https://en.wikipedia.org/wiki/Database_normalization

Although this data may be less easy to read at a glance, it is easier for machines to interpret and can be validated more easily. It can also be transformed into many different views. For example, it is easy to generate cross tabulations of the data to present it similar to the original TJN view.

However, the final measure in the example above (Jurisdictions of NFFEs through which assets are held) does not fit so well. This is a one-to-many relationship. This kind of data is harder to fit into a table. For this reason, standards often use tree structures. XML (eXtensible Markup Language) and JSON are both tree data structures.

In XML, the first and last rows of data above may be represented as:

```
<report>
  <jurisdiction>AR</jurisdiction>
  <start>20130302</start>
  <end>20150301</end>
  <measures>
    <measure code="aggregate_value">
        <value currency="USD">10000000</value>
    </measure>
    <measure code="jurisdictions_of_NFEEs">
        <country code="KY"/>
        <country code="BR"/>
    </measure>
  </measures>
</report>
```

This is called a tree structure because (in the example above) the 'report' element has some common variables, which don't need to be repeated, and it then 'branches' out into separate measures, each with their own variables.

Standards such as XBRL (eXtensible Business Reporting Language) build upon XML by setting out a number of common ways in which measures, dimensions and values should be represented and structured in these data trees. This can lead to stronger validation of data, and the ability to reuse common building blocks from different kinds of business reporting.

Notice, however, that in the tree structure, countries feature at two different levels: at tie near to the root and at other times on a branch as a country code. To find all times a country was mentioned would require knowing where to look in the data structure. Graph data structures address this by modeling data as a set of relationships between entities and properties of those entities.

To allow data from different sources to be joined together, linked data graph structures use URIs (web links) to identify entities, relationships and their properties. This allows computers to navigate data across the web to discover context, providing the data publishers have correctly configured their systems to serve up linked data.

It is possible to convert these tree or graph representations back into tabular format (i.e. to view in a spreadsheet application) although if every datum is included there could be considerable duplication. Often tabular representations will instead provide a summary interpretation.

The transparency platform: focus on data

The FTC transparency platform focuses on Beneficial Ownership, Country-by-Country Reporting, and Automatic Exchange of Financial Information. Each of these creates different open data standardization opportunities and challenges.

Beneficial Ownership: who controls it?[8]

Finding out who the natural persons are who ultimately control or profits from a company is essential in tackling crime and corruption. Yet, right now, those who want to hide their identity can use complex corporate structures, anonymous shell corporations and secrecy jurisdictions to make identification almost impossible.

Registers of beneficial ownership seek to address this. By requiring limited liability companies to report on who their beneficial owners are[9], they close to gaps that can be used to launder money, evade tax or engage in illicit activity.

The European Union has agreed the 4th Anti-Money Laundering Directive to create (restricted access) national level registers of beneficial ownership information through the Union, and the UK has committed to provide a full public register of structured beneficial ownership data, due for first release in mid-2016. In May 2015, Ukraine implemented a requirement for all legal entities to file information on their Beneficial Owners with the State Registrar, and this information has been made accessible in unstructured formats for lookup on a company-by--company basis.

Norway, Denmark and Argentina also have taken steps towards or have established BO registers[10], with Denmark now publishing a register of direct shareholders[11]. Pilots collecting beneficial ownership information have also taken place as part of the Extractives Industry Transparency Initiative, although few of these have yet generated substantial public data[12]. There is also discussion in the UK concerning the creation of a register of beneficial owners in land and property transactions, and the World Bank's 2015 Procurement Framework commits to explore the collection and publication of BO data on entities participating in bank financed projects[13].

There are a number of different ways to assess beneficial ownership, including through voting control, rights to profits and shareholding, and the thresholds for disclosure vary between contexts. In some cases only information on shareholding of 25% or greater is collected, whereas in others, a 5% threshold is set. In EITI pilots, the Kyrgyz Republic has set no threshold for BO by Politically Exposed Persons (PEPs), and the Liberia Multi-Stakeholder Group has recommended that the definition of BO be extended to cover owners of major contractors to an extractives project. In some systems, where no one holds over 5%, then the top-5, or top-10 owners are to be listed.

....

8 Financial Transparency Coalition. (2015). Beneficial Ownership overview (web page)
9 Tax Justice Network. (2015). Financial Secrecy Index Indicator 3: Recorded Company Ownership
10 Global Financial Integrity (2015). Norway Latest Country to Adopt Public Registry of Beneficial Ownership (blog post)
11 See http://datahub.virk.dk/dataset/systemtilsystemadgangtilcvrdata (dataset documentation)
12 EITI Beneficial Ownership Pilots: https://eiti.org/pilotprojectbeneficialownership (reports)
13 World Bank. (2015). Procurement in World Bank Investment Project Financing: Phase Ii: the New Procurement Framework

Key challenges in establishing an open data standard for beneficial ownership information include:

- Uniquely identifying[14] the companies involved in reporting;
- Uniquely identifying specific individual company owners;
- Modeling different dimensions of ownership, benefit and control including shareholding, rights to profit, and voting rights[15];

Country-by-Country Reporting: where is profit made?[16,17]

It is impossible to understand the actions of a transnational corporation in a single country when their financial reports are aggregated on a continental or global basis. Country-by-country reporting would make the activities of consolidated corporate groups more transparent at the scale of democratic jurisdictions.

In particular country-by-country reporting will make apparent cases when corporations use transfer pricing to shift profits to tax havens and costs to high tax countries. This information will help legislators to understand how to change the law to prevent tax sheltering. The information will also guide public interest groups, journalists and consumers to exert pressure on companies not to avoid tax and provide evidence in cases of tax evasion.

Country-by-country reporting could also provide a basis for introducing unitary taxation where companies are taxed according to some measure of their economic activity in a country, not on the basis of legal entities like holding companies.

There are several country-by-country reporting frameworks:

- The Extractive Industries Transparency Initiative (EITI);
- The Dodd Frank Wall Street Reform and Consumer Protection Act (the Dodd Frank Act);
- The EU Accounting and Transparency Directives[18];
- The EU Capital Requirements Directive IV (CRD IV); and
- OECD/ G20 Guidance on Transfer Pricing and Country-by-Country Reporting.

Key challenges in establishing an open data standard for country-by-country reporting include:

- Uniquely identifying the companies involved in reporting;
- Representing the subsidiary-parent networks of ownership as well as joint ventures;
- Incorporating accounting and financial reporting standards to ensure consistency of definitions;

---

14 *That is to say, there is a single, unambiguous reference a number or code that identifies each individual or company*

15 *In particular, there are early indications that countries may release data on ownership shares in bands, reporting with different, noncomparable thresholds. For example, the UK plan to release shareholding ranges from 25%-49.99%, 50%-74.99% and 75%-100%, whereas in Denmark the current bands for direct shareholding data cut across these, for example, including a 66.66%-89.99% band.*

16 *Financial Transparency Coalition. (2015).* Country-by-Country Reporting overview *(web page)*

17 *Tax Justice Network. (2015).* Country-by-Country Reporting overview *(web page)*

18 *EU. (2013).* Directive 2013/34/EU of the European Parliament

- Ensuring reporting can be global, and not limited to a few countries: without a standard that can be used for global reporting, taxable profits could simply migrate to "reporting havens;"
- Setting out what should be public as open data, and what should be restricted to tax authority access only[19].

Automatic Exchange of Information: where is wealth held?[20]

Holding assets overseas allows citizens and corporations to evade tax domestically. If a government wishes to know about the assets and accounts held by citizens in foreign jurisdictions, they must request this information on a case-by-case basis. This process is prohibitively expensive and also entails circular references: the government must know whom they are looking for and where they are holding the money before they are even able to request that very information. It is estimated[21] that the burden of tax losses falls disproportionately on developing nations.

The core idea of automatic exchange is that a country will provide financial information about foreigners to the account holder's home country government at regular intervals.

The G20 and OECD have drafted a Standard for Automatic Exchange of Financial Information in Tax Matters[22]. The standard requires reciprocal exchange between states. It is believed[23] that the relative cost of providing this information will be disproportionately high for developing countries (who may have fewer resources available for tax administrations). This is particularly inequitable if the net movement of wealth is from developing countries to developed ones.

The G20/OECD standard includes:

- A Common Reporting Standard (CRS) that describes the due diligence rules for financial institutions to follow to collect and report information;
- A Model Competent Authority Agreement which provides the legal basis for exchange;
- Commentaries which help to illustrate and interpret the CRS and CAA and
- Guidance on technical solutions, including an XML schema and standards for data safeguards and confidentiality, transmission and encryption;

The Standard builds upon the earlier FATCA laws enacted by the US requiring foreign financial institutions to report information about US citizen to the IRS. Indeed the CRS schema is virtually identical to the FATCA schema in terms of structure and content. Financial institutions dealing with the USA are compelled to comply with FATCA by a 30% withholding tax imposed by the IRS. CRS on the other hand, has no pecuniary or other form of enforcement.

....

19 OECD recommends that country reports remain confidential and are only seen by tax authorities. By contrast, a consultation of large businesses undertaken by Christian Aid, found that "only three companies (6% of those responding) suggested that a concern about commercially sensitive information justified opposition to legislation on public CBCR (four others mentioned it as a concern preventing voluntary reporting)". See Christian Aid. (2015). Country-by-Country Reporting: A survey of FTSE100 companies' views

20 Financial Transparency Coalition. (2105). Automatic Tax Information Exchange (web page)

21 ibid.

22 OECD. (2015). Automatic Exchange (web page)

23 Financial Transparency Coalition. (2105). Automatic Tax Information Exchange (web page)

The CRS requires that the information exchanged be kept confidential and that safeguards are in place to prevent unauthorised access or disclosure.

Given the confidentiality of individual taxpayer information that may be exchanged through automatic information exchange, the questions we must consider for this pillar of the transparency platform are slightly different, namely:

- Is the proposed standard likely to be implementable by developing countries;
- Are there approaches from open data standardisation, which could offer a route to more effectively implementation of automatic exchange of information;
- Are there any datasets related to automatic exchange of information which should be made available as open data, and if so, what standards should be considered for these.

Open data and the transparency platform

The table below outlines the core open data opportunities related to each of the transparency platforms pillars, and how these might be made available.

Transparency Platform Focus	Opportunities for Open Data
Beneficial ownership	Structured registers of beneficial ownership - published by company registration agencies, or other government bodies.
Country-by-Country Reporting	Structured company reports - published by a country's tax authority or company registration agency, and/or by companies themselves.
Automatic Exchange of Financial Information	Aggregate statistics about: a. The process of automatic information exchange; b. Income and deposits reported through automatic information exchange. Bulk data about individuals is inappropriate for release as open data. However, common data standards remain important to make AEoI effective.

Many different groups stand to benefit from the proactive publication of these datasets. In the next section we consider the needs of four core user groups: business, government, journalists and civil society and individuals, with a view if identifying some of the core requirements for data they may have.

Data users and user needs

Understanding how different standards affect the ability of different users and groups to achieve their goals is vital to assessing whether proposed standards deliver against FTC goals, and assessing the political feasibility of different options.

Open data advocacy has often drawn upon a 'big tent' approach[24], seeking to show how open data can benefit a wide range of different users, and supporting coalition building to advocate for data supply. This has proven broadly successful when it comes to the securing the release of data in other domains, although it is important to be aware that different user groups may have different specific needs when it comes to how data is supplied, and which particular data points are included in disclosures[25] .

These use cases draw on interviews and desk research looking at existing data use practice.

Business

Companies currently seek out beneficial ownership information as part of their due diligence when entering into commercial relationships and contracts with other firms. This information is costly to obtain, often incomplete, and results from 'black box' systems that may not provide detailed provenance or source information. There are often restrictions on how the data available can be reused, although the most sophisticated companies want to combine the data they obtain with their own internal data.

In a paper which identifies 15 business use cases for BO data, the B20 group note that *"For companies, there is a growing business need to develop a risk-based approach to prevent money laundering and other financial crimes. This means developing an approach that does not simply look to tick boxes but instead, properly identifies, understands and assesses a company's risk exposure to financial crime. This is imperative because not knowing who truly sits behind the corporate entity you're doing business with can pose significant risk – be it financial, reputational, or to society as a whole"[26].*

Investors also have a particular interest in understanding the global revenues and company operations that may be revealed through country-by-country reporting. They can use this data in order to make better risk assessments and valuations. Investors currently draw upon stock market filings for much of their information. The Eumedion foundation, for example, as representative of institutional investors' interests, believe that "improved disclosures should also help investors make their own assessment of how aggressive and sustainable the tax policy of the company is, and may help identify risks related to taxes" and that "more transparency… will protect the interests of all stakeholders, the reputation of the enterprise, and brand value" leading to "an increase in trust from society and investors"[27].

---

24 Weinstein, J., & Goldstein, J. (2012). The Benefits of a Big Tent: Opening Up Government in Developing Countries. *UCLA Law Review Discourse, 38(2012), 38–48.*

25 Gray, J., & Davies, T. (2015). Fighting Phantom Firms in the UK : From Opening Up Datasets to Reshaping Data Infrastructures?

26 B20. (2015). Putting Beneficial Ownership Transparency into Practice.

27 http://eumedion.nl/en/public/knowledgenetwork/consultations/201509responseectaxtransparencyconsultation.pdf

Government

Financial transparency offers governments the opportunity to increase tax revenues by pursuing individuals or companies who are evading tax and by making wider changes to policy and legislation that prevent tax sheltering.

Government agencies, most notably tax authorities, have an interest in more efficient and cost effective data exchange. Lowering the cost and complexity of data exchange is of particular interest to developing country governments. For example, an overworked developing country tax inspector, without access to the budgets or credit cards needed to subscribe to existing online intelligence sources or to buy company information from another nation's corporate registry, needs access to free and easy to use websites that work in a low-bandwidth and low-technology environment. Right now, regulators, tax inspectors and law enforcement have to use slow processes to request information on tax affairs or beneficial ownership, and unless appropriate treaties are in place, information is not disclosed. When information is provided, it will often be in the form of documents, useful only in the context of a single investigation.

If data is provided in a common standard then there will be economies of scale around tooling. That is to say, the equivalent software (e.g. for storage or analysis) does not need to be written for each and every tax jurisdiction. Developing countries would not necessarily need to develop their own software. Furthermore, revenue authorities working across multiple countries would not need to maintain and operate multiple systems.

Journalists & civil society

Journalists and civil society groups often have similar needs: engaging in investigative activity to uncover illicit, corrupt or otherwise indefensible activity. Many investigations are based on searching for information on specific companies and individuals, using advanced search features rather than carrying out large-scale network analysis across big data. Current investigations are frequently frustrated by the high cost of data, and limited range of analytical tools. Journalists may be particularly interested in linking beneficial ownership data to data on Politically Exposed Persons (PEPs).

If data is published in a machine-readable format then performing calculations on single companies will be much easier and calculations comparing multiple companies - that might otherwise be prohibitively expensive - will become possible. Experience from the use of OpenCorporates data to track elite capture of revenue Myanmar's multi-billion dollar jade industry, has demonstrated how open data can be used to strengthen investigations, and can provide a persistent store of data even when official records and services go offline.[28]

However, it is important to recognise that for many countries, activists are struggling to get access to basic data on country budgets and government expenditure. Data on beneficial ownership or company taxation is even less common. In our discussions with participants at the FTC conference in Jakarta, it was clear that few were thinking as far ahead as the kinds of datasets considered in this report.

....

28 Hussain, H. (2015). How open company data was used to uncover the powerful elite benefiting from Myanmar's multibillion dollar jade industry. (White paper)

Individuals

The data covered by Automatic Exchange, and, to an extent, Beneficial Ownership, is information which may be of direct use to the individuals it describes. Whilst the financial information covered by Automatic Exchange is not intended to be public in its disaggregated individual-level form, strong data standards for this data could provide benefits to individuals: particularly those with complex and cross-border income and tax affairs.

Standards here could enable the creation of 'MiData'[29] or 'Project VRM'[30] services, that let individuals get their data off their banks and financial institutions in a structured way, and that they can reuse within intermediary services to more easily prepare tax returns and understand tax liabilities. This highlights an important opportunity to encourage the creation of customer-centric as well as regulatory-based standards for financial data capture and exchange.

Adversarial use cases

It is important to be aware that it may be possible to use open data on BO, CbCR and AEoI in ways that work against the goals of the Fiscal Transparency Coalition.

We do not provide an in-depth assessment of adversarial use cases in this report. However, it is important to consider how data could be abused to create:

- **Threats to privacy** through adversarial data users triangulating data to reveal private facts about specific individuals, or engaging in large-scale analysis of data in ways that lead to illegitimate interference in the autonomy of individuals and communities;[31]
- **Market manipulation through incorrect data.** For example, if voluntary reported data is relied upon in setting prices or making financial decisions, then there may be incentives for actors to misreport data.
- **Identification of tax loopholes.** With a large corpus of data available, it could be possible for a user to find patterns that help point to new tax avoidance and evasion strategies.

We recommend that in developing advocacy around data standards, further work is done to explore potential adversarial use cases in more depth, and to consider any relevant policy and technology responses to mitigate their impact.

---

29 *MiData, involves the idea that consumers should be able to access machinereadable data that companies hold about them. Examples include the Blue Button, which allows Americans to download their own health records, and the government backed MiData projects in the UK which provide bank and phone company customers with the ability to download their own records in standard formats.*

30 *VRM (Vendor Relationship Management) suggests that the traditional CRM (Customer Relationship Management) model ought to be inverted with consumers holding their own data and having control of how it is shared with vendors (via "fourth party" intermediaries who work for the consumer, not the vendor).*

31 *Solove, D. J. (2005).* A Taxonomy of Privacy. *University of Pennsylvania Law Review, 154(3), 477.*

Towards gold standards

Standards come in many shapes and sizes. 'Gold standards' relating to BO, CbCR, and AEoI would contribute to meeting the core FTC transparency platform objectives, and the needs of a wide range of user groups, whilst also offering a technically robust approach to open data sharing.

This section presents a range of issues to consider when evaluating open data standards, with a particular focus on criteria that would govern choice of an ideal 'gold standard'. In practice, the most feasible standardisation options may not deliver on all of these features, involving pragmatic choices based on priority user needs and understanding the capacities of data publishers in different countries.

Features of gold standards

Global interoperability
It will be much easier to follow the global movement of finance if we have global standards. A single common standard is more efficient to operate, whereas fragmented and conflicting standards create costs for both data producers and users. Where legal and practical differences between countries limit the scope for exactly identical approaches to publication in every context, data should still be produced in interoperable ways: with common crosswalks and mappings maintained between standards. Furthermore, unless the scope of a standard is global, there will remain opportunities for tax evasion and money laundering.

Unique Identifiers
Common identifiers are vital for joining up data. Without them, small differences in the way a name is written, or differences between datasets in the reported address of a company (for example), can lead to matches between datasets being missed, and dramatically higher costs to build an integrated picture of a company's activities across datasets. The FTC transparency platforms requires approaches to identify at least:

- Tax Jurisdictions
  Companies and other legal entities (e.g. trusts)
- Government entities
- Individuals

'Gold standard' identifiers should be unique, authoritative and persistent. That is, each identifier should pick out one and only one entity, and it should always refer to that same entity.[32]

Common base data types
Even simple components of a dataset like numbers and dates are written down differently around the world risking confusion when merging data files. Agreeing on existing standards

....

32 *A recognised 'gold standard' for identifiers on the web is the use of Uniform Resource Identifiers (URIs). Commonly recognised URIs include things like the ISBN number for books, or the Document Object Identifier (DOI) used for referring to academic publications, as well as the URL, or web address. These are issued through a distributed, hierarchical structure which delegates control, and ensures two people cannot issue the same identifier. When URLs are used as identifiers, the owner of that URL can provide additional information about the thing identified when the URL is dereferenced (looked up).*

for basic data types greatly aids interoperability. For example, agreeing to use of YYYY-MM-DD date formats, and agreeing the use of '.' as the decimal separator in numbers.

Common encoding

A globally interoperable data standard needs to be able to cope with data in many different languages and character sets. In the past, systems working in different languages used a wide range of character encodings. Now, encodings such as UTF-8 offer a globally usable common character set: but unless standards specify use of this, even the first stages of data exchange can become difficult and error prone processes.

In theory, adopting UTF-8 is relatively straightforward. It is simple enough to specify UTF-8 encoding when saving a file.[33] Problems typically arise when people edit and save these files without paying attention to the encoding. The default chosen by commonly used applications (e.g. Microsoft Excel) tends to machine and locale-specific (e.g. ISO/IEC 88591 "Latin1"). Because most encodings have the same first 128 characters that cover unaccented English characters, this unintentional corruption is not apparent to many users. Indeed many software developers overlook this critical issue.[34] Thus it is important that standards make this expectation explicit.

Internationalisation

Data typically needs to be written in a single language and read in multiple languages. This is achieved by distinguishing the machine-readable references from human-readable labels.[35] The machine-readable references will be written in a single language so that they may be unambiguously interpreted. The human-readable labels may then be provided in any language. Translating the data is then a case of translating the labels. In the case of field names and code lists, the same set of labels may be used across all records.

Internationalisation (often abbreviated to i18n) is indeed a far deeper challenge than that posed by translation alone. Different "locales" may have different rules for number and date formatting, pluralisation, punctuation text direction and sort ordering. The Unicode Common Locale Data Repository[36] project aims to support a standardised response to catering to these diverse requirements.

Accessible documentation

Standard documentation is often written solely for an expert audience, in ways that exclude many potential users. A standard should be accompanied by clear and user-focussed documentation. This documentation may need to be translated into multiple languages.

Models for privacy, anonymization & redaction

Open data is generally defined to exclude personal data - recognising the right of individuals to privacy, and to control information about themselves (a control that is not possible when

---

33 *In web and email servers a charset attribute may be set on the ContentType header. In HTML a http equiv meta tag is provided for that header (so that the server doesn't need to know about every page it serves). In XML there is an encoding attribute.*

34 *Joel Spolsky. (2003).* The Absolute Minimum Every Software Developer Absolutely, Positively Must Know About Unicode and Character Sets (No Excuses!)

35 *In XML, for example, the xml:lang attribute allows the author to specify the country and language code in accordance with IETF BCP 47 (consisting of RFC 4646 and 4647).*

36 CLDR_Unicode Common Locale Data Repository *(website)*

data is under an open license)[37]. However, in the case of public records, such as the name, address and date of birth of directors on a company register, the picture may be blurred: privacy standards, statutes, and cultural views on whether this class of information is public or private vary between countries.

There may be also cases where full open disclosure of data cannot be made. In these cases, clear policies are needed to govern when data should be anonymised or redacted, and to also ensure anonymisation and redaction is not abused. The International Aid Transparency Initiative, for example, asks all publishers to develop a clear exclusions policy and highlights justified reasons for restricting the detail of disclosures.[38]

Feasible data validation

Validation is a process of checking that data is fit for a given purpose. It allows:

- The creator of the data to ensure that they have entered the data correctly preventing mistakes from the outset
- The data publisher to confirm the integrity of what they release and to ensure that the pipeline they are using to create the data is working
- The data consumer to confirm that the data will be suitable for a given processing step or transformation, providing a warning as to likely problems

Validation comprises a set of rules that describe what valid looks like, and a process by which the data is evaluated in terms of those rules. Different standards languages afford the possibility of different validation tool chains. Given the range of organisations that may need to prepare data that meet BO,CbCR and AEOI standards, the existence of robust validation tools will be important to ensure data quality.

Verifiability

Digital files can, by their nature, be easily modified by anyone with a copy. In recognition that there may be incentives for certain parties to 'cheat' and modify data prior to, or after, publication, it is important for a 'gold standard' to incorporate measures to prevent this.

A range of technologies exists to ensure the authenticity of documents.

Public Key Cryptography allows a file to be digitally signed in such as way as to offer a guarantee of its creator, and that it has not been tampered with by any other party.

Providing assurance that a file has not been modified since it was created, even by it's creator, is more - complex and either involves having some trusted party who can keep copies of timestamps and digital signatures of files - or exploring the use of public ledgers, such as block chain technologies.[39]

....

37 Scassa, T. (2014). Privacy and Open Government. Future Internet, 6(2), 397–413. doi:10.3390/fi6020397 and O'Hara, K. (2011). Transparent government, not transparent citizens: a report on privacy and transparency for the Cabinet Office.

38 IATI. (2014). Exclusions Policy (website)

39 Block chains (the technology that underpins virtual currency BitCoin) rely upon the distributed control of computing power to create a consensus record of some set of facts about the world. Data, such as the digital signature of a file, can be encoded into a block chain, along with a timestamp, embedding this in a public record in such a way that someone would have to gain control of the majority of the computing power in the blockchain network to alter the record.

Immutability

Immutable data cannot be changed. This restriction improves the quality of data by making it more reliable. If data is mutable data (able to be changed) then it is harder to reason about it. We don't know whether the copy, which we have, is up to date. We might not know for what periods of time the data is valid (as this implicit dimension is often missing). Immutable data by contrast can be held in multiple places concurrently and (because of the inherent restrictions) is more likely to include data describing its temporal relevance or applicability. Further immutable data provides us with a historical record and a current record whereas mutable data can only provide the latter.

In the case of company identifiers for example this means that the system should hold a single immutable reference for each company and group of companies. When the company joins or leaves a group the merger or divestment event should be modeled explicitly rather than having records being changed. Circumstances change, but the historical record is permanent.

Balancing expressiveness and simplicity for publishers and users

The adoption of standards is ultimately a question of who bears the cost of data integration. A robustly implemented standard demands a one-off investment from data publishers, and a minor ongoing cost of maintaining compliance, whilst lowering the costs on data users who no longer have to interpret and integrate different datasets. This supports a publish-once, use everywhere approach to data.

However, finding the right level of standardisation is a balancing act. If standards are over-complex, the costs for both publishers and users can be prohibitive. If they are too simple, they gloss over important jurisdictional differences, or the needs of different user groups, leading to some users being excluded from the benefits of the standard.

A particular focus for the FTC is on standards, which support data to be produced and used by developing country actors. The 2nd edition of the Open Data Barometer, reviewing open data readiness across 86 countries, indicates that low-income countries have much lower levels of readiness to create and use open data across government, civil society and private sector[40]. There are no general rules as to which standards will be easier or more complex to implement in resource constrained contexts. However, user testing of any standards in developing country contexts is vital to understand whether or not they will create prohibitive barriers to data production and use. It may be important to develop models for the progressive adoption of a standard, with a basic implementation option that has a low barrier to entry, and more advanced features made optional.

Licensing

Open standards should be clearly licensed, giving users a royalty-free right to represent data using the standard and to build tools and analysis that draws upon standard schemas and documentation. They should not include proprietary content.

Ambiguity about the licensing of standards, identifiers or data can inhibit their reuse. In particular, many of the most productive uses of open data involve combining different datasets and linking public and private data. Terms of use that prohibit this kind of derivative data being shared limit the utility of open data. Non-commercial license terms (which prohibit profiting

---

40 Web Foundation. (2015). Open Data Barometer: 2nd Edition.

from a dataset) can also limit the emergence of sustainable intermediary ecosystems around data. For this reason, both these limitations are not allowed in an open license.

Governance
To meet the needs of all the different user groups we consider, standards should be well governed, and with processes that are open to input from a range of stakeholders.

Many standard governance bodies operate on a paid membership model, where members control the governance process. This can lead to processes dominated by private sector parties, or with limited representation of developing country governments.

Measuring against the gold standard: key questions

In the table below we set out a series of key questions that can be asked of any standard in considering how close it comes to a 'gold standard'.

Feature	Key questions
Global interoperability	Is this standard global, or could it be extended to be applicable globally? Are the basic structures of the standard tied to a specific national or regional use case?
Unique identifiers	Does this standard provide a clear method to uniquely identify jurisdictions, companies and individuals covered by the data? Are the identifiers used non-proprietary? Are the identifiers shared by other standards to enable joined up data?
Common data types	Does the standard make use of existing ISO standards for dates, times, currencies and other core building blocks?
Common encoding	Does the standard specify use of unicode?
Internationalisation	Does the standard have a clear approach to allow multilingual data? Are key codelists available in translation?
Accessible documentation	Is there freely available and well-maintained documentation for the standard? Is the documentation available in multiple languages?
Feasible data validation	Does the standard set out clear policy and technology approaches to validate data? Are there ways to assess the compliance and quality of the data? Is the bar for valid data set too high (preventing less technically advanced countries from producing data), or too low (allowing poor quality data to be provided)?
Verifiability	Is there a mechanism to prevent manipulation of published data?
Immutability	Does the standard recommend providing immutable data? Or does the standard allow publishers to overwrite previously provided data?
Expressiveness & simplicity	Do key publishers and users find they can represent all the key information they want within the standard? How complicated do publishers and users feel it will be to work with data from the standard?
Licensing	Is the standard and/ or data available under an open license?
Governance	How will updates to the standard be managed? Is there an open process for contributing?

In addition to considering the standards through which data will be represented, it is also important to assess the processes and initiatives that will secure commitments from publishers to provide data, that will validate the quality of data, and that will aggregate data together for reuse. The political capital, technical skill and accessibility of these initiatives is likely to have significant impacts upon the quality of an open data ecosystem around each of the FTC transparency platform areas.

Existing standardisation

There are extensive standardisation efforts relating to the building blocks of the FTC transparency platform, although limited evidence of unified and comprehensive standards being adopted at present for each of the pillars.

Standardisation efforts are focused in a number of different areas:

- **Data collection**: setting out how companies and financial institutions should produce data related to CbCR and AEOI;
- **Data transfer:** setting out how data should be exchanged between competent authorities (e.g. between tax agencies in AEOI);
- **Data publication:** setting out how information should be made publicly available.

Decisions about data collection or transfer standards have implications for data publication standards. Ultimately data can only be published if it was collected, and it may be relevant to lobby for data to be published directly by the organisations collecting and submitting it, alongside securing publication of information by the authorities that aggregate it.

Standards for financial concepts, entity identifiers and reporting frameworks exist at varying levels of maturity. Most have emerged over the last 15 years, and a number of standardisation efforts were created in response to the global financial crisis. Most are still under active development, predominantly guided by membership-body governance structures. Our interviews pointed overall to a general phenomena of minimal open standardisation around data collection and transfer in the financial sector, with many organisations used to building on legacy systems rather than open standards. Standards for data publication are even less well established. This may be starting to change with the growth of the Financial Technology (FinTech) industry bringing modern software development practices and ideas from the world of Web startups into financial services, although this does not necessarily equate to open standardisation[41].

When it comes to the specific areas of the FTC platform, work on data transfer standards for AEOI implementation and CbCR are in their early stages, with limited work so far on publication standards. With the first bulk data on BO not anticipated until next year, and limited requirements for data transfer between authorities in the case of BO, there is also limited evidence of established standards. However, there are starting points for standardisation in each of these areas, and we consider the opportunities and challenges of available approaches.

This section starts by setting out the standardisation of the crosscutting building blocks of standardisation in terms of organisational identifiers, individual identifiers, financial taxonomies, reporting frameworks and common data elements, before turning to standards for the specific pillars. Regardless of whether or not comprehensive standards can be identified over the short or medium-term, advocacy on adoption of specific approaches to each building block can proceed, as allowing divergence of approach in these areas will greatly frustrate future standardisation efforts.

....

41 *Standardisation does not necessarily mean open standardisation. Many of the most successful Web 2.0 startups (e.g. Facebook) have succeeded by creating their own proprietary defacto standards, and exploiting these to create barriers to entry, any carefully managing the 'openness' of their platforms in order to ensure they retain a high degree of centralised control.*

Crosscutting elements of standardisation

Organisational identifiers

Organisations frequently change names and address. The need for unique identifiers for companies was recognised long ago when company registers were set up and starting issuing company numbers. The need for unique public identifiers for government entities has only been more recently recognised, as data-driven projects face the challenges of integrating data from multiple sources and struggle with reconciling different phrasings of the same department or agencies name. When it comes to company ownership, a large number of other legal vehicles can also be in the frame, and not all are subject to clear registration schemes that provide them with identifiers. For example, trusts and partnerships may have the status of legal persons in some jurisdictions without being officially registered or assigned any official identifier.

Key candidate standardisation opportunities for organisation identification include:
- The IATI Organisation Identifier methodology;
- The **Global Legal Entity Identifier (GLEI)**;
- **GLEI Business Card information**.

The **International Aid Transparency Initiative (IATI) Organisation Identifier methodology**[42] is premised upon the re-use of existing registered identifiers of organisations. It provides a codelist of known organisation registration agencies which can be used as a prefix to the identifier assigned by that agency to generate a globally unique ID. For example, for a UK company registered by Companies House, and with the company number '0123456' would get the prefix 'GB-COH' (Country Code + COH for Companies House) to be added to its number, to generate a globally unique organisation identifier: 'GB-COH-0123456'. This methodology has also been adopted by the Open Contracting Data Standard (OCDS)[43]. However, at present the codelist of registration agencies has limited coverage, and many countries do not provide open data from their corporate registries, limiting the amount of contextual information that can be looked up from an identifier.

To be adopted as part of a 'gold standard', the IATI Organisation Identifier methodology would need to move onto a more stable governance footing: managed by a consortium of standards groups, rather than a single open data standard.

The main alternative to the IATI Organisation Identifier methodology takes a different approach. Instead of reusing existing issued identifiers, it looks for each reporting entity to register a new identifier.

The **Global Legal Entity Identifier (GLEI)** emerged as part of regulatory responses to the 20072008 global financial crisis, with the recognition that counterparty identification was a major regulatory gap. First developed in the US Office of Financial Research (created by the Dodd-Frank Act of 2010), the idea for standard legal identifiers for financial firms was quickly internationalised through the G20 initiated Financial Stability Board (FSB). Announced in 2012, the LEI system has been developed through a layered governance structure, with the Global Legal Entity Identifier Foundation (GLEIF) and it's Central Operating Unit responsible for setting data standards and managing a network of Local Operating Units (LOUs) which issue LEIs to

....

42 IATI. (2014). Organisation Identifiers (website)
43 Open Contracting Partnership. (2014). Open Contracting Data Standard (website)

firms and verify their data. This is overseen by a Regulatory Oversight Committee (ROC), who approve the strategic direction of the LEI. Firms have to pay a fee to apply for an LEI, and at present, only legal persons participating in financial markets are eligible for LEIs.

The current LEI data standard, determining the information captured against each identifier, is focused on provision of 'Level 1' business card information, providing an XML schema for describing the basic name and address details of firms, along with a single related firm field. The standardised data from LOUs is gathered daily and published in bulk at https://www.gleif.org/.

At present, around 400,000 GLEIs have been issued, suggesting the LEI is unlikely to have broad enough coverage to identify all the entities of interest in the FTC transparency platform. However, the **Business Card Information** captured in the Common Data Format of the GLEI can provide a model for capturing enough to increase the change of disambiguating information on entities, and matching together data on entities from different sources. The Common Data Format[44] captures structured data on legal name, other known names, registered address, other addresses and business register ID (amongst other things).

Individual identifiers

Public identifiers for individuals are vital in BO registers, as well as in databases of Politically Exposed Persons, and robust (though not necessarily public) identifiers of individuals are required in automatic exchange of financial information. However, privacy concerns and varying cultural attitudes create considerable barriers to a clear standard for assigning identifiers to individuals. Unlike company identifiers, where the goal may be to identify a single scheme which leads to one, and only one, authoritative identifier per legal entity, when it comes to identifying individuals, there is generally no single authority. Instead, identifying information, which can support matching of records, is required.

There are three main options available for identifying individuals within a dataset:

1. **Providing adequate personally identifying information -** such as name, date of birth and address details. A name on it's own is not enough, nor is birth year and month in most cases due to the number of matches for these variables.

   There is significant complexity here due to the way names are translated and entered into databases[45]. Recent efforts to construct an aggregated Persons of Interest dataset[46] rely on capturing and storing extensive Also Known As (AKA) data, and alternative spellings of names, providing data users with as much information as possible to allow a best effort at finding matches.

2. **Reusing an existing identifier, such as Taxpayer Identification Number (TIN).** Even though an individual with business in multiple jurisdictions may have multiple such identifiers, the consistent use of such identifiers can help reconcile datasets. For example, if automatic information exchange reports are made in multiple years, the incoming identifier only needs to be matched to the internal identifier by the receiving government the first time it is encountered. Whether or not a TIN, Passport Number

---

44 Global Legal Entity Identifier Foundation. (2015). Common Data File Format (website)

45 For example, many data entry systems have input boxes for 'first name' and 'last name', yet when these are filled in by individuals where the cultural practice is to put family name first, or with dual family names (e.g. Spanish communities), the data that gets populated will not follow this pattern, and it will be unclear how different users have interpreted these boxes.

46 Friedrich Lindenberg. (2015). Open Names (webpage)

or other personal identifier, should be made part of the public record or not varies between cultures and jurisdictions.

3. **Assigning individuals a unique identifier in each system.** For example, when a company register creates its own internal identifier for a shareholder and uses this in published information about companies that individual has a shareholding in. Again, consistent use of such identifiers assists reconciliation of data about specific individuals.

In general, current approaches focus on (1) although in some cases internal identifiers may be exposed in structured data.

It is beyond the scope of this report to consider the different privacy and policy considerations involved in publishing personally identifying information in different jurisdictions. This is an area in need of further investigation.

Financial reporting taxonomy

Almost all of the substantive data we are concerned about will come from financial reports. The accounting terms used in these reports are subject to interpretation (e.g. a measure of profit may or may not take into account taxes, depreciation, interest, or dividends). In order to refer to these line items consistently we need a standard taxonomy, backed by clear and detailed definitions.

In order to supply the global applicability of a standard it is also important that a taxonomy can be made available in a range of relevant languages. For this reason, well-structured taxonomies may provide codes which are then described and documented in a number of languages.

Many countries have their own accounting standards, but increasingly the **International Financial Reporting Standards** (IFRS) are being adopted, or accepted, by jurisdictions. The IFRS Foundation report that *"14 of the G20 jurisdictions have adopted IFRSs for all or most companies in their public capital markets[47]"* and that *"more than 125 countries either require or allow the use of IFRS for all or most publicly accountable entities when preparing financial statements[48]"*.

An XBRL taxonomy (see below) for IFRS is available, although evidence from the IFRS foundation suggests there is currently very limited developing country adoption of electronic filing using this[49], with the only reported developing country case of adoption being a voluntary filing option due for introduction in Malaysia by early 2016.

It is notable that the USA does not use IFRS, instead relying upon its own Generally Accepted Accounting Principles (US-GAAP). The IFRS Foundation report that *"The SEC permits but does not require its foreign private issuers to use IFRS as issued by the IASB in preparing the issuer's financial statements. The SEC does not permit its domestic issuers to use IFRS in preparing their financial statements; rather, it requires them to use US GAAP[50]"*

---

47 IFRS. (2014). Analysis of the G20 IFRS profiles
48 IFRS. (2014). IFRS Filing Profile Project
49 Ibid.
50 IFRS. (2015). Jurisdictional Profile: United States of America

It is beyond the scope of this report to recommend a full approach to advocacy regarding financial taxonomies when it comes to standardisation on FTC transparency platform priorities. However, we note that to generate globally comparable data at source will require use of taxonomies that agree on how core terms are defined (and thus how calculations to generate figures such as profit and sales should be conducted), or that have robust crosswalks to support mapping from one taxonomy to another. The burden of reporting on a global standard set of definitions, rather than those in use domestically, falls primarily upon the businesses responsible for reporting. Though, ultimately, consistent global definitions will reduce the costs for business operating across different jurisdictions.

Reporting frameworks

A number of different approaches exist for the exchange of reporting information, ranging from spreadsheet templates through to rich schema and taxonomies for representing complex information in machine-to-machine communication.

Spreadsheet templates

Early work on Country-by-Country Reporting has focused on simple spreadsheet templates for reports, whether in the Extractives Industry Transparency Initiative, or OECD model template for CbCR[51].

Well-designed spreadsheet templates can incorporate data validation, ensuring dates, codes and numbers are entered accurately, and in ways that enable the data to be easily converted into structured data which can be queried. Spreadsheet templates also support direct use of the data, without it being imported into database systems, greatly aiding the accessibility of the information.

However, poorly designed spreadsheet templates often make it very difficult to aggregate together data from different sources, and often struggle to capture more complex data.

Statistical Data Cube and SDMX

OnLine Analytical Processing (OLAP) Data Cube models provide a way of representing statistical data as a combination of dimensions and measures[52]. The example of standardised tabular data in the box-out above was an example of a data cube model. Here, each row represents a single measure, and then all the different dimensions by which that measure is classified.

The Statistical Data and Metadata eXchange (SDMX) standard[53] provides a suite of XML schemas and data structure definitions for representing cube modeled statistical data, including data for national accounts, trade statistics and balance of payments[54]. A linked data ontology to model SDMX as linked data also exists in the form of the RDF Data Cube Vocabulary[55].

---

51 OECD. (2014). Guidance on Transfer Pricing Documentation and Country-by-Country Reporting
52 Wikipedia: OLAP Cubes
53 SDMX. (2015). What is SDMX?
54 SDMX. (2015). SDMX Implementations at the International Level, by Statistical Domain
55 W3C. (2014). The RDF Data Cube Vocabulary W3C Recommendation 16 January 2014

Custom XML

The eXtensible Markup Language (XML) allows the expression of complex data structures, and supports the validation of data. It is commonly used for machine-to-machine data exchange, although it is more difficult for users to read directly. Some reporting schemes, such as the Common Reporting Standard Schema for Automatic Exchange of Information define their own XML schema.

XBRL

The XBRL International consortium describes XBRL as a "language in which reporting terms can be authoritatively defined. Those terms can then be used to uniquely represent the contents of financial statements or other kinds of compliance, performance and business reports. XBRL lets reporting information move between organisations rapidly, accurately and digitally"[56].

The language provides common identifiers (or "bar codes for reporting") to allow financial reports and records of transactions to be tagged. XBRL provides taxonomies for concepts from a variety of accounting and risk domains "including [International Financial Reporting Standards] IFRS, US [Generally accepted accounting principles] GAAP, Chinese GAAP, Japanese GAAP, and the European Basel III bank reporting framework CRD IV and insurance reporting framework, Solvency II amongst many others"[57].

XBRL is used in 50 countries, being adopted by, among others, the U.S. Federal Deposit Insurance Corporation, the Committee of European Banking Supervisors (CEBS), Securities and Exchange Commission (SEC) in the United States, the United Kingdom's HM Revenue and Customs (HMRC), Companies House in Singapore and the Ministry of Corporate Affairs (MCA) of India. Mature tools exist for working with XBRL, with considerable support in accounting software and data management platforms.

It is also important to recognise that there are two parts of any XBRL standardisation: the use of XBRL as the reporting language, and then selection of a taxonomy against which reports will be made. In many instances, even developed economies only have limited XBRL implementation. For example, in the UK, only accounts prepared to the UK Generally Accepted Accounting Practice (GAAP) are filed via XBRL. Accounts prepared to IFRS are not currently included, and revised accounts are also not included in the open XBRL data that UK Companies House makes available, suggesting these are not captured either[58].

Universal Business Language

Universal Business Language (UBL) is a long-established XML Schema for interchange of business objects. It provides agreed terms for a wide range of 'business objects', primarily oriented towards order fulfillment (tenders, ordering, shipping, invoicing, payment) to be used *"in generic procurement and transportation contexts"*[59].

....

56 XBRL International. (2015). An Introduction to XBRL The basics of XBRL - for business and accounting professionals.
57 XBRL International. (2015). Getting Started for Developers - Introductory information aimed at professional IT developers and programmers that would like to utilise XBRL.
58 Companies House. (2015). Free Accounts Data Product FAQs
59 OASIS. (2013). Universal Business Language Version 2.1

UBL is a well-established standard, with the first version published in 2004, and the current version approved for use in procurement within the EU, and submitted to the International Standards Organisation (ISO).

Whilst not oriented towards financial data, the payments components of UBL have been used by the UK within schema to implement the EU Accounting and Transparency Directives.

FIBO

The Financial Industry Business Ontology (FIBO) provides an extensive collection of terms for describing key features of the financial landscape. Developed by the Object Management Group (OMG) (a membership based standards body), in collaboration with the Enterprise Data Management Council (EDMC) (a nonprofit trade association), it has so far published a 'Foundations' ontology, and an ontology on Business Entities (FIBO-BE) which describes different forms of relationship between firms and people. The project aims to ultimately provide vocabularies for detailed descriptions of ownership and control, partnerships, trusts, corporations and other legal entities.

FIBO is linked data ontology, designed to support representation of data using the RDF (Resource Descriptor Framework) format. Linked data has proved prohibitively complex in many prior open data pilots in developed countries, and the associated tooling and documentation is relatively inaccessible. As such, there are high barriers to adoption of RDF as a reporting format. However, the terms and data structures from an ontology can be used without using the RDF format.

Due to the early stage of FIBO development, adoption of the standards is unclear, with limited evidence of government adoption. The paid membership structure of EDMC also means we were not able to access full documentation on the current status of FIBO development, although documents from OMG suggest an active process of development and there appears to be substantial participation from financial firms in the development of the ontology.

At present, FIBO terms appear to be defined in English language only, and defer to the *Barrons Dictionary of Finance and Investment Terms*[60] for definitions, which may affect their global applicability over the short-to-medium term.

Common encoding and data elements

Encoding errors, differences in country names and confusion over the interpretation of dates are just some of the common challenges when working with data. Although in each of these areas well established standards exist, the best of which are based on a robust process that has considered how to handle differences around the world, they are often ignored.

Well developed reporting frameworks such as XBRL and FIBO will already define common encoding and data elements to be used, but when standards are proposed based on flat-file and spreadsheet structures, this is less commonly the case. We do not provide, an exhaustive list of relevant standards here, but to note some of the most important:

---

60 Downs and Goodman. (2014). Barrons Dictionary of Finance and Investment Terms

- UTF8[61] is a unicode standard for character encoding design, ensuring characters from all alphabets can be represented in stored text.
- ISO[62] 8601 provides a standards for date and time formats[63], based on YYYY-MM-DD, and including time zone offset information for date-times.
- ISO 3166 provides 2digit country codes and subdivisions[64] maintained by a group of ten organisations worldwide, and ensuring no ambiguity over the country being referred to;
- ISO 4217 currency codes[65] add to ISO 3166 two letter codes a single letter corresponding to the first letter of the currency name.
- ISO 31 and 800001:2009 provide standards for representing numbers
- RFC[66] 4180 Common Format and MIME Type for Comma-Separated Values[67]

These standards have specific technical implementations. For example, country codes are implemented as an XML Schema by Code Synthesis Tools CC[68] and currency codes in the OASIS Universal Business Language Schema[69].

It is important to ensure emerging standards build upon these well-established standards.

Domain specific standards

Beneficial ownership

No country is yet publishing a full beneficial ownership register as structured open data, and there are no clear standards emerging.

The OpenCorporates Open Company Data Index[70] highlights that practice around the world on publishing company registers is very varied, with different levels of detail provided publicly, and many only accessible through data scraping, rather than through bulk open data.

In Denmark, where a register of shareholders is now available, this is published using a custom JSON structure[71]. The structure indicates the owner, and their share of ownership in 10 non-equal bands[72].

---

61 See http://www.utf-8.com/

62 The International Organisation for Standards (ISO) develops and promotes (the voluntary adoption of) worldwide proprietary, industrial and commercial standards.

63 ISO 8601

64 ISO Country Codes

65 ISO Currency Codes

66 The Internet Engineering Taskforce (IETF) aims to "make the Internet work better by producing high quality, relevant technical documents that influence the way people design, use, and manage the Internet". The IETF develop and publish Requests for Comments (RFC) which document suggestions (although they do not specify mandatory internet standards as such, they tend to enjoy widespread adoption).

67 IETF RFC 4180

68 Code Synthesis. (nd). XML Schema Standard Type Library

69 OASIS. (2015). OASIS Universal Business Language

70 Open Corporates. (2015). Open Company Data Index (website). See also: Open Knowledge. (2015). Open Data Index

71 VIRK. (2015). System-til-system adgang til CVR data

72 Thresholds at 0.05, 0.10, 0.15, 0.20, 0.25, 0.33, 0.50, 0.67, 0.90, and 1.00

In the UK, the beneficial ownership register, which is currently under construction, will capture five different kinds of control[73], and percentage shares in three bands from a 25% threshold, as well as a range of statements by companies with respect to the quality of data provided (e.g. to indicate whether they are still seeking further disclosures from shareholders). The information, initially submitted to Companies House annually, will have to include the details and dates of changes over the previous period.

In Ukraine, beneficial ownership information is currently captured through free-text reporting in the company register, although there are plans for more structured data capture to be introduced in future.

Most beneficial ownership regulations to date have looked at disclosure of private company ownership, as separate reporting regimes exist for public listed companies generally through stock market filings. In investigation of US SEC data, we could not find a standard structure to filings related to BO.

Starting points for standardisation

Promising starting points for the development of a gold standard include:

- Extension or adaptation of the **Global Legal Entity Identifier Level 2 data model** (currently under development).

  A consultation recently took place to explore standards for 'Level 2' relationship data, with the proposed options focusing on identifying immediate and ultimate parent entities, with the scope restricted to corporate parents, and with accounting consolidation definitions used to judge the nature of any relationship.

  Chan and Milne (2013)[74] note that *"understanding corporate hierarchies...will be valuable byproducts of the LEI system"* but caution that *"these must not be the first priority in its initial design, lest 'overreach' discourages adoption and weakens the entire project"*. This is evidenced in the cautious approach of the LEI ROC in consulting over Level 2 data.

  Given the scope of the LEI only covers legal entities participating in financial markets, it does not offer a full solution to either corporate hierarchy or beneficial ownership data standardisation challenges. However, it will increasingly offer relevant and open-data backed identifiers for many organisations involved in BO hierarchies, including those located in secrecy jurisdictions[75], and the data structures used for Level 2 data may be flexible enough to be reused for BO, offering opportunities to align a global BO data registry with the LEI without overloading the LEI adoption process itself.

  This depends upon the extent to which the final schema from the Level 2 consultation process is flexible to inclusion of new relationships beyond those covered by accounting principles of control.

---

73 Shares; Voting Rights; Ability to Appoint the Board; Otherwise exercise significant influence or control; Owning a company through a trust or similar firm or similar structure.

74 Ka Kei Chan and Alistair Milne. (2013). The Global Legal Entity Identifier System: Will It Deliver?. *Economics Discussion Paper Series. Loughborough University.*

75 For example, the LEI database currently includes 100s of Cayman Island registered firms.

- Work by OpenCorporates on the prototype BO database WhoControlsIt[76], and on incorporating ownership structures into its own database. The open learning from this work with real-world data provides a good starting point for a standardisation process.

    Building an open source reference implementation of a BO register also offers an opportunity to demonstrate how governments may be able to implement data collection and publication at relatively low cost, providing they have existing mechanisms for authenticating those reporting information.

Gold standard goals

A gold standard for BO information will need to contain the following elements:

1. Identifiers for organisations, including private and listed companies;
2. Identifiers for individuals;
3. A taxonomy of types of ownership and control;
4. A set of recommended bands for capturing/publishing ownership shares[77];
5. A framework for recording changes over time;
6. A framework for capturing whether ownership is direct or via declared or undeclared intermediate parties;
7. A framework for capturing the provenance of information;

In addition, reports from EITI pilots of beneficial ownership call for a building block relating to:

8. The identification of Politically Exposed Persons (PEPs).

1, 2 and 8 are crosscutting standardisation challenges. 3-6 are domain specific, and may require further work at a policy level before full data standardisation can be effectively pursued. 6 combines policy and technical considerations.

Country-by-Country Reporting

Each of the current Country-by-Country reporting frameworks currently in operation propose their own standards, with varying degrees of technical complexity and precision.

Frameworks

EITI
The EITI International Secretariat has published a standard reporting format for summary data[78]. Those reporting under the initiative are required to publish a report and then provide data using the template to the secretariat via email. The data is then used to populate the global EITI data repository and made available online.

---

76 See http://alpha.whocontrolsit.com/

77 A gold standard should recommend the publication of actual percentages, but, given there are already legal frameworks in place (e.g. UK) which only collect data in bands, it will need to allow a fall back to banding. As such, it is strategically important to ensure interoperability of data that an approach to banding is agreed. For example, agreeing bands, when used should be at 5% intervals.

78 EITI. (2015). Summary Data Template

The template is implemented in Excel with some validation and fields to handle URLs or references to sections of the report. This method is simple enough to understand but is far from an ideal solution. The data must be reported twice (in the report proper and the template). Validation in Excel spreadsheets is not rigorous and extracting data is error-prone (indeed the template requires that reporters alter its structure to suit their requirements). Submission by email is hard to automate effectively - manual intervention will be required (e.g. to distinguish initial submission and corrections etc). Several ISO standards have been adopted, however, and a structured template is certainly an improvement on PDF reports.

OECD Guidance on Transfer Pricing and Country-by-Country Reporting

The 2014 OECD Guidance on Transfer Pricing and Country-by-Country Reporting[79] provides a series of model templates for Country-by-Country reports from companies, setting out fields that could be included, but stops short of technical schema for how the data should be represented. The implication from this report was that companies might use this structure in tables within PDF reports, rather than encouraging machine-readable data publication. It would presumably be down to each state to translate this into any national standards for machine-readable data collection from companies.

However, the 2015 OECD Action 13: Country-by-country Reporting Implementation Package[80] suggests that *"it is intended that an XML Schema and a related User Guide will be developed with a view to accommodating the electronic exchange of CbC Reports"* within the framework of Multilateral Competent Authority Agreements (the agreements which allow the exchange between state bodies of CbCR data). This emerging data transfer standard will inevitably have implications for the technical implementation of machine-readable data collection from companies, and for future data publication.

Section 1504 of The Dodd-Frank Act

Section 1504 of the Dodd-Frank Act requires extractive industries companies that report to the SEC to provide detailed information about their payments to any government, anywhere in the world, on a country-by-country and a project by project basis. Section 1504 establishes an "interactive data standard for the information included in the annual report of a resource extraction issuer". These reports should include "electronic tags that identify, for any payments made by a resource extraction issuer to a foreign government or the Federal Government: (I) the total amounts of the payments, by category; (II) the currency used to make the payments; (III) the financial period in which the payments were made; (IV) the business segment of the resource extraction issuer that made the payments; (V) the government that received the payments, and the country in which the government is located; (VI) the project of the resource extraction issuer to which the payments relate; and (VII) such other information as the Commission may determine is necessary or appropriate in the public interest or for the protection of investors".

Because the Securities and Exchange Commission requires[81] reports to be filed in XBRL format, Section 1504's requirement that the information be provided "interactive data standard for the information included in the annual report of a resource extraction issuer" means that the information to be disclosed pursuant to Section 1504 must also be provided in XBRL. The

79 OECD. (2014). Guidance on Transfer Pricing and Country-by-Country Reporting
80 OECD. (2015) Action 13: County-by-county Reporting Implementation Package
81 SEC. (2015). XBRL Reporting Portal (website)

statutory language allows the format of data collected to change over time as the standards used in SEC reporting change, however, as recommended later in the policy section.

Since, for the most part, the Dodd Frank Act deals with financial reform, there are other standards that, while not directly relevant to illicit financial flows, may be of tangential interest.

The Financial products Markup Language[82] (FpML) is an XML schema for electronic dealing and processing of over-the-counter derivatives. It establishes the industry protocol for sharing information on, and dealing in, financial derivatives and structured products.

Unique Swap Identifier are devised to identify swap transactions[83]. "The first characters of each USI will consist of a unique code that identifies the registered entity creating the USI given to the registered entity by the Commission [...]. The remaining characters of the USI will consist of a code created by the registered entity that must be unique with respect to all other USI's created by that registered entity".

The EU Accounting and Transparency Directives
Directive 2013/34/EU, in paragraph 39 states that "The Member States are strongly

Name of the MNE group: Fiscal year Concerned:										
Tax Jurisdiction	Revenues			Profit (Loss) Before Income Tax	Income Tax Paid (on cash basis)	Income Tax Accrued- Current Year	Stated Capital	Accumulated Earnings	# of Employees	Tangible Assets other than case and cash equivalents
	Unrelated Party	Related Party	Total							

Guidance on transfer pricing documentation and country-by-country Reporting ©OECD 2014

encouraged to develop electronic publication systems that allow undertakings to file accounting data, including statutory financial statements, only once and in a form that allows multiple users to access and use the data easily. With regard to the reporting of financial statements, the Commission is encouraged to explore means for a harmonised electronic format. Such systems should, however, not be burdensome to small and medium-sized undertakings". That is to say, the directive does not provide a single pan European data standard.

In the UK, for example, where the directive has been transposed into The Reports on Payments to Governments Regulations 2014[84], there is a draft extractive reports XML schema[85] published by Companies House. The schema builds upon Universal Business Language (UBL) and ISO

82 Financial Products Markup Language *(website)*
83 CFTC. (2012). Unique Swap Identifier (USI) Data Standard
84 UK Government. (2014). The Reports on Payments to Governments Regulations 2014
85 UK Government. (2015). Extractive industries reporting: draft schema and step by step guide

codes but not XBRL.

The EU Capital Requirements Directive IV (CRD IV)

CRD IV mandates the collection of increasing amounts of banking capital and exposure data throughout the countries of the European Union. An XBRL taxonomy, Basel III XBRL, is provided for this reporting.

Starting points for standardisation

Whilst a number of countries have established XBRL-based reporting routes for submitting machine-readable country-by-country reporting data, globally, XBRL adoption remains patchy, and has a high technical barrier to both data publication and use.

By contrast, both the EITI and OECD have focussed on providing reporting templates, albeit templates that stop short of supporting robust data validation, and that will not necessarily generate interoperable data.

A focus on providing and testing machine-readable versions of these templates, with clear data validation, offers a plausible starting point towards greater standardisation, not least because it provides a relatively straightforward ask of companies, and results in individual data files which are easier to analyse, whilst holding out the prospect of aggregating together data submitted by multiple companies.

Associated with this, work on defining key terms, and aligning the vocabularies used in reporting, is likely to be important to support globally interoperable data.

Gold standard goals

A gold standard for Country-by-Country Reporting should support:

1. Direct publication by companies of their reports;
2. Publication via country tax authorities;
3. Use of common coding for jurisdictions;
4. Use of common codes and encoding for currencies and financial figures;
5. Provision of company identifiers;
6. Structured data on the relationship of each company to the group;
7. Structured data about the financial period covered by the report;
8. Clear definitions of all key terms;
9. Accessible presentation of summary data in spreadsheet form;

Automatic Exchange of Financial Information

The OECD has a well-developed process in setting standards for transfer of information between tax authorities under the Common Reporting Standard (CRS), building on a schema from FATCA. Standards for the publication of aggregated information are yet to be developed.

The OECD Common Reporting Standard (CRS) for Automatic Exchange of Information requires that data be exchanged between competent authorities using XML documents compliant with a specified schema[86]. There are no prescriptions over how the financial institutions should collect information and transfer it to tax authorities although it is suggested that the same schema be used to obviate the need to reformat the data. The required schema is derived from the FATCA schema[87] in terms of structure and content meaning that tax authorities and financial institutions reporting and exchanging information under FATCA should not need to make significant additional investment. However, it is not clear whether the governance of FATCA and CRS schemas are aligned, such that they could diverge in future.

The CRS also requires that minimum standards of for secure (encrypted) transmission to prevent disclosure to unauthorised persons and to ensure the integrity of the data:

- In terms of encryption, the standard suggests the use of public key cryptography and the AES-256 cipher (Advanced Encryption Standard)
- In term of transmission, the standard recommends (browser-based) HTTP (Hypertext Transfer Protocol) transport with TLS v1.1 (Transport Layer Security), SFTP (Secure File Transfer Protocol), or secure e-mail
- In terms of operations, the ISO 27000 family of Information Security Management Systems are recommended

There are several initiatives underway to ensure that developing countries have the capacity to implement the CRS. The G20 has given a mandate to the Global Forum on Transparency and Exchange of Information for Tax Purposes to help developing countries identify needs for technical assistance and capacity building, working together with the OECD Task Force on Tax and Development, the World Bank Group and others[88]. There is a Roadmap for Developing Country Participation[89] describing how developing countries can overcome obstacles and pilot projects to partner lower capacity countries with more experienced peers. It is also hoped that the staged implementation of the Standard will ease adoption.

Since the information being exchanged relates to private individuals and entities, the CRS is designed to ensure confidentiality rather the open publication of data. In the general case there is a widely recognised cultural norm affording individuals the right to privacy. This is by no means universal however, in Norway, for example, is it possible to request data about the taxes paid by specific individuals (although this is not published by default). Moreover, in the case of Politically Exposed Persons, there is likely wide support for open publication as it may help to guard against bribery and corruption.

---

86 OECD. (2015). CRS Schema
87 IRS. (2015). FATCA XML Schemas and Business Rules for Form 8966
88 OECD. (2015). Automatic Exchange of Financial Account Information - Background Information Brief
89 OECD. (2014). Automatic Exchange of Information: A Roadmap for Developing Country Participation

Even if all individual reports remain confidential there is still scope for the publication of aggregations as open data. These statistics could include measures of the total flows between countries and the rates of compliance. The Tax Justice Network has proposed[90] that aggregate statistics be collated by each pairwise combination of account holder country and financial institution country with a further breakdown by type of financial institution. For each grouping, there would be summary statistics for a) aggregate and median values of account balances and annual income and b) count of accounts and account holders. For the first year of AEOI this would also include the number of new and closed accounts so as to capture avoidance schemes.

The Tax Justice Network have put forward spreadsheet templates for this aggregate publication, but further work is needed to develop a robust machine-readable standard for such disclosures.

There may also be value in the open publication of linkages between Tax Identification Numbers assigned to the same individual in different jurisdictions. This would allow tax authorities to reap economies of scale when reconciling identifiers. Although these identifiers themselves will be pseudonymous (i.e. they use codes rather than names) it may be possible to de-anonymise identifiers. This can be done by using other metadata, such as dates of changes or sets of tax jurisdictions, to identify individuals by their information footprint. There is a trade-off between efficiency of data analysis and protection of individual privacy. Moreover this concern applies to aggregate statistics too where the breakdowns could be disclosive.

Starting points for standardisation

The existing XML schema work from the US Government for FATCA, and from the OECD for the CRS should provide a solid basis for data collection and transfer standards. Engagement with the finalisation and further development of these standards to ensure they adequately capture all key information, and will enable interoperability between jurisdictions is likely to be the most effective way to shape the possible supply of data.

With a focus on the publication of aggregate data, analysis of the data collection and transfer standards should explore through pilot work the extent to which the CRS and FATCA schemas will enable robust production of aggregates. This may require a focus on the validation rules and processes that will be applied to judge whether data is of adequate quality when it is submitted.

Tools to round-trip data from simple tabular formats into the structured XML required by CRS and FATCA may be relevant in supporting developing country engagement with data production and use by supporting at least basic data preparation, validation, publication and use in spreadsheet tools[91].

In order to support the production of robust aggregate data, a data cube based data model

------

90 Knobel & Meinzer. (2014). "The end of bank secrecy"? Bridging the gap to effective automatic information exchange: An Evaluation of OECD's Common Reporting Standard (CRS) and its alternatives

91 For example, the Convert, Validate and Explore (CoVE) tool developed by Open Data Services takes data published in the Open Contracting Data Standard and converts it to spreadsheet formats whilst carrying out data validation. This approach has proven useful to keep programme and management staff engaged with the production of data as they can use spradsheet software they are familiar with to review data produced. If data production is left solely in the hands of technology team there is more room for errors in the data to be missed, as it may be structurally valid, but semantically incorrect.

could be developed, drawing upon terms and definitions articulated in the CRS and FATCA XML schema, and setting out XML or spreadsheet templates for row-by-row reporting of each required measure. Simple spreadsheet pivot charts, or more advanced data validation and presentation applications can then be used to present this information.

Gold standard goals

In considering gold standards for AEOI, it is important to focus on:

1. A clear vocabulary of terms of each dimensions and measure in the data - with harmonised definitions;

2. Use of common coding for jurisdictions;

3. Use of common codes and encoding for currencies and financial figures;

4. Structured data about the financial period covered by the report;

5. Clear processes for data quality validation;

6. Clear processes for transforming granular data into aggregates;

7. Clear policies on data privacy, and approaches to aggregation and anonymisation;

Costs of implementation

In developing advocacy for countries to adopt standards for collection, transfer and open publication of key data, it is important to be aware of how costs and benefits are distributed. In many cases, developing countries have less developed e-government infrastructures to build upon, but, in implementing new systems can build in open standards from the start.

Other things being equal, a standard that is easier and cheaper to implement should be preferred. The preceding discussion focuses on quality considerations. In this chapter, we consider the role of costs in the choice of a "gold standard". In particular we are seeking to provide insight as to whether the standards being proposed in the developed countries are appropriate for the institutional environments in developing countries.

Balancing the burden and benefits

Producing and using data involves costs. Standards affect how these costs are distributed.

When users encounter data in a format and structure, which fits well with the purpose they had in mind for the data, their costs of accessing, interrogating and understanding the data are lower. When the data is either more complex than they are prepared for, or does not include key elements they need, their costs of data use will be higher. The degree of complexity in terms of technical format and data structure that users can accommodate will depend upon their technical skills and the market for intermediary services that package data appropriately for different users. As a rule of thumb, business users are most likely to be well served by intermediary services and to have access to the skills to use data effectively, with developing country governments, media and NGOs less equipped.

Data publishers face similar variation in their costs: when a data structure fits well with how they manage data internally, and when they have already invested in the skills and tools to generate data in a particular format their additional costs of adopting a standard are lower than when they have to purchase and learn new technologies, or when they have to adjust their business processes to generate data that will fit into the standard being used[92].

In some areas, the benefits of standardisation accrue both to publishers and users. In other cases, standardisation operates primarily for the benefit of data users. For example:

- Without unique company identifiers, data users face the cost of reconciling and matching company lists. To be able to publish data with reliable unique company identifiers a country may need to re-engineer various business processes, data collection forms and databases. However, at the end of this process the government will also benefit from being able to manage its own information more effectively.

  This investment is best made by the publishing country, but brings a benefit both to users and to the government making the reforms.

....

92 *In developing an advocacy strategy it can important here to also consider which ministries or agencies of government will bear the costs of adaptation: as these may not be the ministries with the strongest interest in the benefits that standardisation will bring.*

- By contrast, there may be a more limited direct benefit to a country of adopting a global accounting taxonomy; particularly if the standard imposes a taxonomy which is not widely used in the region the country is in. Either users must bear the costs of mapping between taxonomies, or the country must invest in mapping its reporting to a common taxonomy, without necessarily seeing a strong return.

Understanding how the costs and benefits of individual elements of standardisation play out is important to understand the incentives of different actors. It also highlights the importance of considering the sunk investments that countries already have in certain technologies and practices.

One further important consideration in assessing the costs of benefits of standardisation relates to the timeline for a return on investment. For some standardisation actions (again, organisation identifiers is a good example here) benefits can be immediately felt. For others, there is a network effect: the benefit only occurs when a critical mass of other data publishers adopt the standard.

The cost of not standardising

While there are costs to implementing standards, there are also costs that arise from having not implemented a standard.

Creating a custom schema is often easier than fitting into an existing one but this simply defers or displaces the burden of integration. Each tax authority may develop their own formats internally but the multinational companies must still incur the costs of reporting against each one.

Not agreeing on common standards can also have costs in the long term. The case of character encoding (discussed above) demonstrates this. The ASCII (American Standard Code for Information Interchange) encoding provided 128 unaccented English characters. Most commonly these would be stored in 8-bit bytes, meaning there was space in memory for a further 128 possible characters. Without an agreement on what the remaining characters should represent there was soon dozens of alternatives. The cost of this non-standardisation is difficult to estimate, but it seems reasonable to suggest that it would have been cheaper for everyone had a common standard been agreed from the outset.

Prevalence of Adoption

If a country has decided, and succeeded, in implementing a standard, we may reasonably conclude that it was affordable to do so. As such, we can learn about the relative cost of different standards by considering which countries have implemented them. This analysis is, of course, confounded by the differences between countries in the value that the standard generates, the assets already existing (e.g. technology and skills) and the policy environment. Nonetheless, we believe the comparison to be informative.

We have reviewed case studies on xbrl.org[93] to identify countries who have implemented XBRL or who are pursuing projects in this area[94]. The box plots[95] below demonstrate the comparison. The countries implementing XBRL do tend to have higher Gross Domestic Product and Gross National Income per Capita[96]. XBRL is by no means, however, the preserve of developed nations. Three of the four BRIC economies (with the exception of Russian) are working with XBRL. The list also includes countries with low total GDP, such as Latvia, Malaysia and Panama, and low income per capita, such as Colombia, Indonesia, and Peru. Moreover, XBRL is being pursued in eleven of the top twenty secrecy jurisdictions[97].

---

93 *We also used this list:* https://www.xbrl.org/the-standard/why/who-else-uses-xbrl/

94 *This includes: Australia, Belgium, Brazil, Canada, Chile, China, Colombia, Denmark, Finland, France, Germany, India, Indonesia, Italy, Japan, South Korea, Kuwait, Latvia, Malaysia, Mexico, Netherlands, Panama, Peru, Poland, Singapore, Spain, Sweden, Turkey, United Arab Emirates, United Kingdom, and the United States. Note that this not an exhaustive search.*

95 *The left and right edges of the boxes show the 1st and 3rd quartiles, the line in the middle shows the median (2nd quartile) and the horizontal lines on each side extend to the minimum and maximum values.*

96 *2014 figures from World Bank's World Development Indicators. The difference is statistically significant.*

97 *As ranked in the Tax Justice Network's* Financial Secrecy Index *2015*

Impact Asses...

This section provi... ...plementing policy in each of the three ...

Beneficial Own...

The UK Departme... ...ed an impact assessment[98] estim... ...ip register. The companies registr... ...9k for the IT development of th... ...k pa on-going for the maintenance. Cos... ... cost (familiarisation, identification, coll... ...or information), and £77.7m pa on-goi... ...rns to Companies House".

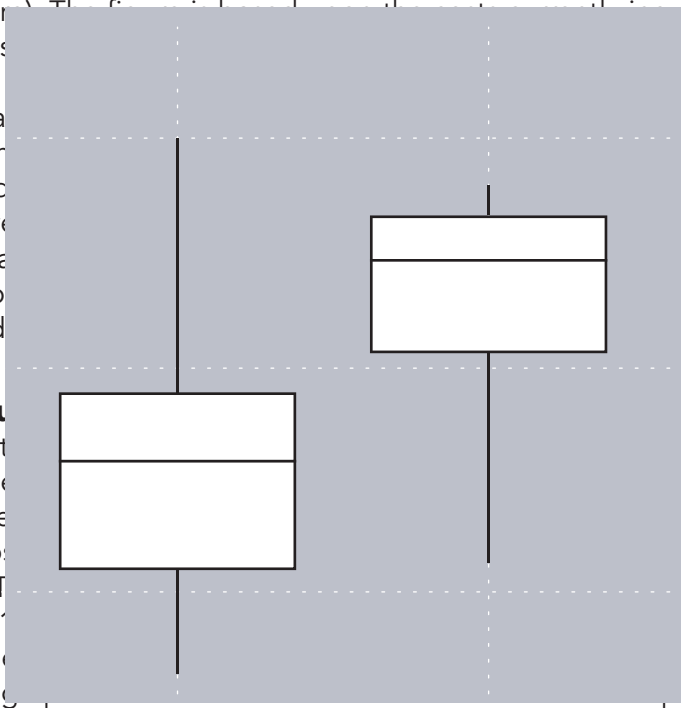The initial costs to... ...dd new fields and tables within the e... ...make the data publically available, and £11.5 to communicate the requirements to corporate directors with inserts sent with existing reminders (this is actually a £23k cost shared between two policies). The top end of the range includes a £39k contingency. Note that all of these costs are operate on the basis of existing company registrar infrastructure i.e. they don't include the cost of establishing a company database and regular communications between registrar and companies.

The ongoing costs to Companies House relate to the additional manual work required to maintain a closed register of beneficial ownership (i.e. in cases where the owner would be at risk from disclosu...). The fig... is based on the costs ... re the i...ured to employ seven people to adminis...

The costs to priva... ...oke quantitative survey (previous research... ...e-off cost per company, £129.70, included... ...taining external advice (£56.30). There we... ...ges and £19.40 for external advice), a... ...0 and for reporting to Companies Ho... ...rived on the basis of companies alread...

Country-by-Cou...

The cost of report... ...costs will be incurred by government in the... ...nalysis of the data. The UK Departme... ...ed an Impact Assessment of co... ...ccounting Directive (2013/34/EU)[100]T... ...g costs for companies will be £11.9m ($1... ...um thereafter across 251 incorporated larg... ...ity. These estimates were reached by scaling... ...nies and an estimate

98 BIS. (2014)/ Transparency & Trust - Enhanced Transparency of Company Beneficial Ownership IA No: BIS BE022

99 John Howell & Co Ltd, Global Witness. (2013). Cost of Beneficial Ownership Declarations extending HM Treasury/DTI. (2002). Regulatory Impact Analysis: Disclosure of Beneficial Ownership Unlisted Companies

100 BIS. (2014). Implementation of Chapter 10 of EU Accounting Directive (2013/34/EU)

Chart 1 (top): Y-axis label "Gross Domestic Product (US Dollars)" with gridline values 10,000,000,000,000; 100,000,000,000; 1,000,000,000. X-axis categories FALSE, TRUE. Chart title "Are pursuing XBRL?" Series "GDP PerCapita"

Chart 2 (bottom): Y-axis label "Gross National Income Capita (Atlas method) (US Dollars)" with gridline values 100,000; 10,000; 1,000. X-axis categories FALSE, TRUE. Chart title "Are pursuing XBRL?" Series "GNI PerCapita"

from the US Securities and Exchange Commission (SEC) consultation into the costs of Dodd Frank compliance.

The BIS assessment also notes that reporters will be able to offset this against the indirect economic benefits likely to accrue to them as a result of more stable operating environments. In particular the assessment notes that Royal Dutch Shell reports that 195,000 barrels of oil a day remain trapped in Nigeria, costing the company nearly $16m.

A study by ICAEW for the European Commission estimates that the costs of preparing International Financial Reporting Standards (IFRS) consolidated returns[101]. The estimates for adoption range from 0.31% of turnover for firms smaller than €500m and 0.05% for those larger than €5b. The recurring costs are estimated to be 0.06% and 0.0008% respectively.

Automatic Exchange of Financial Information
There will be costs to both financial institutions (who prepare submissions on individual and entity account holders) and governmental bodies (responsible for exchanging data).

The evidence from FATCA provides a useful benchmark to estimate the cost of implementing AEoI.

In terms of the investment required by the US Government, the Treasury Inspector General for Tax Administration has estimates[102] the of cost of creating the IRS IT system to support FATCA, the Foreign Financial Institution Registration System, to be $16.6m.

The UK's HMRC estimates[103] that it would itself incur one-off IT and staff project costs of approximately £5m and ongoing annual costs of £1.4m. The aggregate cost to 300,000 UK businesses was estimated to be an initial £2-3b with ongoing £100m - £170m per year. After a USUK Agreement which reduced the reporting scope this estimate was revised down to include 75,000 business, the initial costs were estimated at £0.9b - £1.6b, with an ongoing cost of £50m - £90m a year.

Since the CRS will be common across jurisdictions we might expect economies of scale by comparison with FATCA although this will depend upon how IT systems are procured.

Attempts to estimate the costs to government are complicated because different countries start with different levels of infrastructure. The OECD AEOI compliance ratings[104] provide a measure of government preparedness-serving as a useful proxy, although it doesn't allow us to draw quantitative conclusions.

Factors affecting costs

The estimates provided by economic impact assessments are useful, particularly where we can compare the costs for different aspects within the same study. Each study, however, is based upon a context that is, to a large degree, inseparable. The impact assessments made by the

101 ICAEW. (2014). Document discussed in the Expert Group on the IAS Regulation, meeting 24/10/2014, Agenda item III – Paper 1
102 Treasury Inspector General. (2013). Foreign Account Tax Compliance Act: Improvements Are Needed to Strengthen Systems Development Controls for the Foreign Financial Institution Registration System
103 HMRC. (2013). The International Tax Compliance (United States of America) Regulations 2013
104 OECD. (2015). Global Forum on Transparency and Exchange of Information for Tax Purposes Phase I and Phase II Reviews

UK's Department of Business, Innovation and Skills (BIS) for example, are implicitly founded on the premise that a national registrar of companies exists. Since Companies House has been in operation since the mid 19th Century, it is very difficult to a) imagine what the costs might be without Companies House and b) estimate the costs of establishing and running such an institution for 150 years.

We can identify the factors that affect the costs of implementing open data standards:

- **Institutional framework -** for example: Does a registrar for companies already exist? Does the tax office provide identifiers for individuals? Legal or logistic requirements often mean that governments already have these assets in place.

- **Existing reporting** - XBRL is well established with most large US and UK companies already having to use it for reporting. Those financial institutions that have adopted FATCA will already be most of the way towards supporting CRS.

- **Software and tooling -** open and well-establish standards will have better ecosystem support. Where software vendors and open source projects are already using the standard, it should be cheaper and easier to implement new projects.

- **Scale and scope -** the more reporting that needs doing, the more it will cost shouldn't expand scope without recognising this implication.

This last aspect is considered in more detail in [Appendix 1](#), which sets out three maturity models for assessing steps towards providing open and standardised data.

Funding strategy

In order to understand the cost of open data we must also consider how the implementation is funded.

The simplest scenario would be to fund projects out of general tax revenues. In developing countries, many e-government projects are resourced through donor investments. In these cases, it is important to consider whether both capital and revenue costs are being covered, and how a system will be sustainable beyond the original donor investments.

Another option is to charge fees. In the UK, for example, the Government Trading Funds Act 1973 allowed Crown services to recover their costs by charging fees. Companies House, for example, charge companies for making filing that the companies are legally obliged to make. These 'publisher pays' approaches can ensure that access to the data is provided freely. This is also the approach taken in the Global Legal Entity Identifier, where companies pay an annual subscription fee for their identifier

A third option is the development of Private Public Partnerships, in which companies run key registries in return for certain rights to exploit data in order to raise revenue to fund running the service. These may operate on publisher pays models, but may also adopt a user-pays model of charging for data access. Anecdotal evidence suggests some governments have pursued such approaches due the difficulty of hiring technology development teams on fixed government pay scales.

These private or part-private funding strategies will require less money from tax revenues. They also, however, bear hidden costs. Data exhibits characteristics of what economists call "public

goods"[105]. This means that the social value of data tends to be more than the sum of private benefits it creates. Data that is treated as a market commodity will thus tend to be underused (from the perspective of society).

Summary & conclusions

Costs of implementing the systems and standards required to produce and use high quality open data in each of the FTC transparency platform areas will need to be assessed on a Country-by-Country basis. Developing countries are likely to face some additional costs to reach the levels of data quality and use that can be achieved in countries with well established e-government platforms, although there exists substantial variation in levels of e-government development both in developing and developed countries.

Crucially, it is important to address the cost-benefit analysis of standardisation, and to consider the long-term costs of not standardising, as well as to look critically at the models for funding data collection and transfer to ensure they promote the sustainable supply of open data.

---

105 Data often has high initial cost of creation but negligible marginal costs of re-production, it bears increasing returns to scale, it is difficult to withhold access to data, it may be used by multiple people

Political analysis

Promoting a supply of standardised and open data can take a number of forms, from interventions to influence existing standards processes, through to development of new standards initiatives.

Open data is vital to increased financial transparency. Yet data standards are just one element of strategic efforts to secure accessible, reliable and reusable flows of open data that can be used to address corruption and illicit financial flows. As the previous chapters have identified, there are various components of standardisation relevant to each of the FTC transparency platform areas, yet no comprehensive open data standard initiatives exist which can be given unequivocal backing.

In this chapter we first consider a number of strategic approaches and theories of change related to increasing open data supply. We then consider a number of ongoing political processes which present upcoming opportunities to advance action on the supply and standardisation of open data to bring about greater financial transparency. Finally, we look at the key political opportunities and challenges in relation to each area of the FTC transparency platform.

Open data strategies

Some data is valuable on its own. If you are investigating a specific company, then a Country-by-Country Report from that company is all you need. In other cases, data gains its value through network effects. It is either more valuable the more comparable data is available alongside it, or, in some cases, it is only useful once a complete and comprehensive data from multiple sources is available.

The first stage in developing any strategy based on open data is to understand whether the theory of change for data use relies upon reaching a threshold of data availability, or whether benefits start to accrue as soon as the first bit of data is published.

This leads to the second key question: how with an ecosystem of high quality data be maintained? It is usually not enough for stakeholders to just be using a standard to publish data scattered across their own websites. Unless that data is brought together by some intermediary, the cost of using data from 53 multiple sources will remain high and will limit use. And unless there is a process to check data quality - the standard may be poorly applied.

Open data and transparency projects have responded to these challenges in a number of different ways.

Data scraping and platform building: OpenCorporates
OpenCorporates.com is a privately funded project to build an open database of all the companies in the world. The data is sourced either directly from company registries, or by screen-scraping data (writing scripts which browse through online company directories and extract data). It uses an internal homegrown schema for storing company data, mapping data from each national registry to its internal data model.

The resulting dataset is licensed under the Open Database License (ODbL)[106], with a commercial version available for users who do not want to abide by the 'Share Alike' terms of the license and want to use the data within proprietary products. These commercial services act as a revenue stream to support maintenance of the data, whilst the core open license protects against a 'bait and switch' approach in which the data is later only available for a fee.

The OpenCorporates approach demonstrates the potential to collate a large dataset without a focus on prior standardisation of data at source. However, where governments do not already publish a corporate registry, OpenCorporates cannot collect this. An exploratory prototype by OpenCorporates called 'Who Controls It[107]' points to the potential to build simple tools which capture required data on Beneficial Ownership, intervening with technology and proof-of-concept rather than data standards to show how states should collect data.

Political commitments and technical standards: International Aid Transparency Initiative

The International Aid Transparency Initiative (IATI)[108] launched in 2009 with a declaration calling on governments to provide better forward looking and disaggregated aid information. This political initiative sparked the creation of an XML data standard[109]. Countries are encouraged to sign up to IATI as an initiative, and within this to commit to publication of data according to the IATI standard. Countries are then responsible for adapting their systems to produce data that meets the standard.

A technical secretariat, ultimately overseen by members of IATI, stewards development of the data standard and provides a registry of published data, validation tools and a datastore which aggregates together data which each government publishes on their own website. An independent advocacy group, Publish What You Fund[110] has created an annual 'Aid Transparency Index', ranking countries based on factors including their implementation of the data standard[111]. IATI was also named as an element of the common standard for aid information arising from the 4th High Level Forum on Aid Effectiveness in Busan in 2011. Together these factors has played a significant role in incentivising governments to improve adoption of the data standard, and to increase their data quality.

IATI has the goal of securing 100% coverage of official overseas development aid described using the standard.

User centred standards: Open Contracting Data Standard

Many governments already publish data on tenders and contract awards. Some regional standards exist, such as the European 'Tenders Electronic Daily' XML schemas, and elements of procurement are covered by UBL and other business interoperability standards. However, existing standards were not oriented towards supporting transparency, and meeting the needs of corruption hunters, governments monitoring procurement efficiency and effectiveness and SMEs seeking business with government.

---

106 *Open Data Commons:* Open Database License
107 *See* http://alpha.whocontrolsit.com/
108 *See* http://www.aidtransparency.net/
109 *See* http://iatistandard.org/
110 *See* http://www.publishwhatyoufund.org/
111 *Certain metrics within in the index are measured by looking at data published in a country's IATI XML file, with scores weighted higher when the data is available in the machinereadable standard.*

The Open Contracting Data Standard (OCDS) was developed to fill this particular niche. It offers a standard for interoperability of data, and a framework for assessing data quality. Very little is made mandatory in the standard: but tools and processes are being developed to assess the quality of data produced, and to engage in constructive dialogue with governments about how to improve their data over the short and long-term.

OCDS is based on a model of voluntary standardisation, and can be adopted at the country, city or organisation level. Adoption of the standard has been supported through political advocacy in the Open Government Partnership for countries to adopt the standard as part of wider open contracting reforms, and a focus on Open Contracting within World Bank funding for procurement reforms in client countries.

Mandatory reporting: FATCA and CRS
Where data is collected by a central authority they may mandate the use of specific data standards. This is the case for the US FATCA reports, where financial institutions and host country tax authorities are required to produce data according to a given XML schema, and to report this through a centralised online tool.

Often, the body in charge of the mandate to report has full control of the standard used for reporting. However, there may also be consultations on the development of these standards, creating opportunities to influence the data that is collected, how it is structured, and how it might be shared.

A mandatory standard has a good chance of bringing about comprehensive data in a common format but it often leaves control of the standard, data and decisions about what to release in the hands of a single organisation.

Selecting a strategy
The examples above illustrate that there is a choice to make between a focus on:

- **Creating dataset rather than standards -** by investing effort in mapping from existing published data;
- **Developing standards initiatives -** to combine political commitment to publish new data, technical standards for its publication, and oversight of implementation;
- **Developing user-centred standards -** and working with a range of stakeholders to promote flexible approaches to standard adoption;
- **Influencing mandatory reporting standards -** both engaging at a technical level in the details of vocabularies, schema and policies on openness, and at a policy level to provide a watchdog function on implementation.

This is not an exhaustive canvas of strategies. However, it helps frame to the specific opportunities and challenges relating to each area of the transparency platform.

The policy landscape

Each of the pillars of the FTC transparency platform is subject to ongoing policy processes at global, regional and local levels. Strategic engagement with these processes is needed to track when standards are being proposed, and to push for an 'open by default' approach wherever

possible: avoiding the legislative lock-in of clauses that may prohibit public sharing of BO registers, or information from CbCR and aggregate AEOI statistics. In general, policy processes should set out the principles, but not the details, of standardisation. The inclusion of explicit technical specifications within law can create a barrier to the required evolution of standards, limiting global interoperability. However, principles mandating the use of 'open standards', calling for data to be published under open licenses, and requiring use of globally agreed taxonomies can all be introduced to a process.

In parallel to these domain specific policy processes, there are a range of emerging open data-related policy processes which offer important opportunities for the pursuit of standards, particularly in relation to BO. These include:

- **Open Government Partnership (OGP)[112]**. The OGP is a voluntary association of states pledged to create and implement Open Government Action Plans. There has been a strong focus in the OGP on open data and technological innovation. Initiatives such as Open Contracting and EITI have successfully used the OGP as a platform to secure government commitments to adopt principles and data standards, and to engage national civil society in these processes.

- **The International Open Data Charter (IODC)[113].** Emerging from the G8 Open Data Charter, the IODC was launched alongside the UN General Assembly in September 2015, and sets out a series of principles asking governments to be 'open by default'. It's backers are seeking to sign-up over 40 national and sub-national governments as charter signatories by 2018. A process, led from the President's Office in Mexico, has just begun to develop accompanying technical guidance for the charter, detailing the kinds of datasets that should be disclosed, and the standards that should be used, in specific sectors. The first of these will be an anti-corruption focussed package. The exact status of this guidance is to be defined, but it is anticipated it will at least include BO as a core dataset within a proposed anti-corruption data infrastructure. A consultation[114] on the guidance was launched at a fringe meeting of the 2015 G20 meeting in Turkey.

- **G20 Anti-Corruption Working Group.** The G20 Anti-Corruption working group has published a set of "G20 Anti Corruption Open Data Principles"[115], based on the IODC and describing how *"Open data can help prevent and tackle corruption, accordingly to national law and experiences, by shedding light on government activities, decisions, and expenditures"*. Although the principles do not address issues of corporate transparency directly, under the previous Australian presidency, the G20 committed to "High-Level Principles on Beneficial Ownership Transparency"[116], albeit that stop short of requiring public disclosure.

- **UK Anti-Corruption Summit.** UK Prime Minister David Cameron has committed to holding an anti-corruption summit in Q1/2 2016 in London. This is seeking 'game changer' ideas to feature, and there is a strong possibility a number of these will look at open data, given the UK's strong role within the open data community.

112 Open Government Partnership: http://www.opengovpartnership.org/
113 International Open Data Charter: http://opendatacharter.net/
114 IODC. (2015). Putting the Open Data Charter into practice: Consultation on anti-corruption resource package
115 G20 Anti-Corruption Working Group. (2015). G20 Anti-corruption Open Data Principles
116 G20. (2014). High-Level Principles on Beneficial Ownership Transparency

- **Global Partnership on Sustainable Development Data (GPSDD).** Launched to take forward work on the UN Data Revolution agenda, the Global Partnership describes itself as "global network of governments, NGOs and businesses working together to strengthen the inclusivity, trust and innovation in the way that data is used to address the world's sustainable development efforts"[117]. Although primarily focussed on statistical data to support implementation of the SDGs, it's role as a multi-stakeholder collaboration, and the presence of targets within SDG Goal 16[118] on anti-corruption and illicit financial flows creates a potential space for advocacy on data supply and data standards relating to the FTC transparency platform.

There is strong alignment between the first four of these processes, with many of the same actors heavily involved in each, and anti-corruption clearly on the agenda. The extent to which the GPSDD will ultimately provide space for a focus on SDG Goal 16, governance and anti-corruption remains to be seen.

Working through open data policy processes provides an opportunity for quick-wins on data supply and standardisation, albeit shallow gains - as open data initiatives can generally only unlock data already held by the state, at least over the short term.

Political opportunities for the transparency platform

Alongside wider engagement in building an open data ecosystem for financial transparency, through a watching brief on developments around the Global Legal Entity Identifier in particular, there are specific opportunities and challenges related to each of the FTC transparency platform areas.

Beneficial ownership
Given the early stage of work on BO, and the absence of existing global standards, and opportunity exists to promote the development of an accessible and user-centred standard.

This could act both as a tool to enable interoperable data publication, but also as an advocacy tool to promote best practices in the design of data collection systems for BO.

The existing OpenCorporates platform is well placed to act as a primary public user of this data, filling the role of data aggregator within the emerging ecosystem, and providing demonstration of the benefits from good quality publication. There is also an established industry of firms who would also ultimately benefit from standardised data[119].

Existing monitoring tools, such as the Financial Secrecy Index, Open Data Index and Open Data Barometer, could be used to provide regular monitoring of publication data and implementation of common standards.

An initiative of this form is not a small undertaking, and would need to strike a balance between an agile approach, and building a strong multi-stakeholder support base, potentially

---

117 *Global Partnership for Sustainable Development Data: http://www.data4sdgs.org/who-we-are/*

118 *UN. (2015).* Goal 16: Promote just, peaceful and inclusive societies

119 *Although in the short term, introduction of a standard may appear as a threat to the business models of organisations such as Bureau van Dijk, which potentially rely their substantial investment into creating BO datasets acting as a barrier to new market entrants.*

drawing upon leading governments (e.g. UK, Ukraine, Denmark etc.), CSOs, multilaterals (e.g. World Bank), and private sector engagement.

There are also opportunities to insert explicit commitments to beneficial ownership open data within the Open Government Partnership process, and the International Open Data Charter sector packages, particularly the anti-corruption package work currently underway.

Providing tangible demonstrations of the benefit that BO data brings will be essential to sustain engagement over the long term, particularly as many countries will have to make substantial investments in their data collection systems before BO data is available.

Country-by-Country Reporting
The OECD have committed to developing a XML schema for CbCR. The opportunities to engage with this are unclear, but it is a very relevant advocacy target likely to substantially shape both the details of data that is collected, and to determine data quality by virtue of decisions taken about data validation and quality thresholds.

The OECD standard is focussed on data transfer between authorities, but will likely impact upon the reporting frameworks for companies to submit data to government. There is no guarantee of open data publication at the end of the workflow of companies reporting to governments, and governments transferring data. This highlights an opportunity to work on developing best practice standards for voluntary publication by companies of machinereadable CbCR data as part of their annual reporting.

With project-by-project reporting coming into place for extractives companies, working with extractives industry groups to develop templates for proactive publication by companies of their data may offer one opening. This approach is unlikely to achieve comprehensive data, but does offer an approach to set precedent and to pilot data standards in ways that could inform both the technical details of the OECD process, and advocacy on an 'open by default' approach for CbCR reports.

Automatic Exchange of Financial Information
Development of the FATCA and CRS standards for AEOI is well advanced. Whilst the data gathered through this standards will not be transparent, it is important to scrutinise these standards and the processes surrounding them to ensure that the data that will be collected will be of high quality, and that it can be used to generate aggregate transparency reports.

Concerns about privacy are likely to present a barrier to agreement on disclosures of aggregate information, and so will need to be addressed with a robust and standardised approach. There is a choice to make between focussing on a common technical standard for aggregate statistics disclosures, or focussing on the development of a platform that will bring this data together for analysis.
The most substantial challenge with respect to AEOI is how far the process of data exchange will be accessible to developing countries, both because of their legal frameworks around data privacy, and the technical complexity of handling data. Given programmes of support are in place, external monitoring of this support may be valuable to ensure it is adequately focussed on the needs of developing countries, and that lessons from implementing in developing countries are being captured, learned and fed into future developments.

The governance of the CRS standard is also an area for attention.

Although not currently on the agenda, there may be opportunities to insert a focus on aggregate financial and tax reporting information into discussions on an anti-corruption open data package, or future work on developing the open data ecosystem for anti-corruption.

Summary & conclusions

The open data landscape is rapidly developing. From an early focus on 'raw data now', advocacy organisations are now turning to the creation of data new infrastructures which will provide 'open by default' data for a wide range of users.

Whilst there are clear use cases for open data on the ultimate owners of companies, country-by-country company reporting, and aggregate information on financial and tax affairs of individuals, working towards 'gold standard' open data on each of these will necessarily involve a wide range of interventions. Building blocks for standardisation exist, although gaps also remain, including with respect to widely agreed approaches to company identification.

To take the development of an open data layer of the FTC transparency platform calls for a number of overlapping strategies:

1. **Engaging with open government and open data policy processes** to secure commitments to publish key data, particularly company registers and BO data;

2. **Critically engaging with existing standardisation processes** around CbCR and AEOI, paying close attention to the technical and policy decisions made, the governance of standards, and the engagement of users in testing the accessibility o the models proposed;

3. **Supporting the development of lightweight and prototype standards and data aggregation platforms** particularly for BO and direct publication of CbCR by companies. Such interventions can help shape the debate, providing a demonstration of the value of collecting specific data fields, adopting particular data structures, or pushing for an open-by-default policy with respect to key data.

Whilst we have not been able to recommend specific data standards in most cases, our hope is that the information and conceptual tools outlined in this report can equip member of the FTC open data working group to engage productively in ongoing debates.

Appendices

Appendix 1: Technical maturity models

There is no single technological approach to opening data. Different choices have different cost implications. A maturity model provides a broad categorisation of approaches to distinguish a scale from simpler, cheaper approaches to more advanced and expensive ones.

We present three maturity models: sharing, standardisation, and openness.

The maturity scales presents a trade-off between up-front cost and average costs. In the early stages of a scale advancement makes publishing more efficient, the later stages make consumption more efficient. The ratio of benefits to case in either case depends upon the scale of activity.

Sharing

1. **Post** is the most basic way of sharing data in hard copy or with digital media. While the technical requirements are very low (basic office facilities) but the costs can be relatively high.

2. **Email** is perhaps the simplest online delivery mechanism is to send files via email on a perrequest basis. This requires negligible up-front cost (as email is widely available) but may have a very high marginal costs (as each publication requires manual activity and is only delivered on a one-to-one basis).

3. **Shared Directories** made available online via a web (http[120]) or file (ftp[121]) server is more efficient as it only need be published once and then many people can download the data. Although delivery is cheap for the publisher, discovery is expensive for the consumer (they may have to download everything as review it to find the content they are interested in).

4. **Content-management systems** provide a navigational structure, supplementing the data with metadata such as titles, categories or summaries. Affordances such as menus, indexes, or search allow people to discover content relevant to their interests. Although the user has more power to find the data, they must still undertake a manual process which, again, cannot be scaled-up efficiently.

5. **Application Programming Interfaces (API)** provide a structured means of discovery and delivery to machines. That is to say, an API allows users to build software on top of open data. This allows, for example, automatic updates to process new data as it is released or the flexibility to prepare one visualisation that can be reused to present different data (by changing the data being requested with API parameters).

Standardisation

The main beneficiary of standardisation is the data consumer although this may also be the publisher in cases where they main the data for their own use.

120 Hypertext Transfer Protocol
121 File Transfer Protocol

1. **Proformas/ templates** are the simplest form of standardisation is to agree a basic data model. In excel for example, a template could provide row and column headings to describe the records and attributes or code lists to define permissible values for given fields. These tend to be technologically unsophisticated and are therefore comparatively easy to design and use (at small scales) although they impose hidden costs on interpretation.

2. **Adoption of common data types** to reduce the amount of preparation required by the consumer. Even within the same excel template it is possible to record a date in a multitude of (often incompatible) ways. More advanced templates stipulate the use of e.g. ISO date codes.

3. **Verifiable schema** to ensure data integrity and handle complex validations. Although it is possible to perform data validation within Excel, only simplistic (single value) validations can be enforced. Verifiable schema (such as DSDs and XSDs in the case of XML) can include far more complex rules.

4. **Integration with other schema** for comparability and interoperability. Declaring a schema unilaterally is easier (and thus cheaper) than creating a data model that works with those of third parties. This may be as simple as adopting third-party identifiers or as complex as providing multiple representations of your data in formats compatible with different systems.

Openness
The Five Star deployment scheme describes a maturity curve for open data, the basic tenet being to release early and improve later:

1. **Open Licenses** to release data on the web (in any format). It is possible to make data available without it being open. Restrictive licenses or even just the absence of guarantees to the user are enough to "close" the data, preventing use, modification and sharing. The OKFN maintains a list of licenses compliant with the [Open Definition](#) such as Creative and Open Data Commons licenses.

2. **Machine-readable** data structures that may be understood by a computer (to allow software to extract, transform, and process it). Data that is published in hard copy may be considered open, but it is far from useful. Similarly it may be accessed digitally (e.g. a PDF image of a document) but still not be usable as it cannot be interpreted by a machine.

3. **Non-proprietary formats** that do not require "closed source" software. Even openly-licensed, machine-readable data may be restricted by being encoded in a proprietary format. This means that proprietary software is required to access the data limiting its potential for reuse.

4. **Uniform Resource Identifiers** to denote things, so that people can point at your data. Identifiers allow us to disambiguate between resources, discover new ones and connect to them. Data with identifiers is open for reuse by others.

5. **Links to other data** to provide context. Identifiers allow others to link to your data, the reciprocal step is to link to others' data. Connecting a datum provides context for interpretation. These relationship are what makes data informative. There are positive network effects to data linking with the value of data increasing exponentially as the linked data cloud grows.

Appendix 2: Glossary

AEOI	Automatic Exchange of Financial Information
BO	Beneficial Ownership
CbCR	Country-by-Country Reporting
CSV	Comma Separated Values
EDMC	Enterprise Data Management Council
EITI	Extractives Industry Transparency Initiative
FACTA	Fair and Accurate Credit Transactions Act
FIBO	Financial Industry Business Ontology
FpML	Financial Product Markup Language
FTC	Financial Transparency Coalition
GAAP	Generally Accepted Accounting Principles
GLEI	Global Legal Entity Identifier
GLEIF	Global Legal Entity Identifier Foundation
GTFS	General Transit Feed Specification
IATI	International Aid Transparency Initiative
IETF	Internet Engineering Task Force
IFRS	International Financial Reporting Standards
ISO	International Standards Organisation
JSON	JavaScript Object Notation
OCDS	Open Contracting Data Standard
OECD	Organisation for Economic Cooperative and Development
OMG	Object Management Group
RDF	Resource Descriptor Framework
RFC	Request For Comments
SDMX	Statistical Data and Metadata eXchange
UBL	Universal Business Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USI	Unique Swap Identifier
XBRL	eXtensible Business Reporting Language
XML	eXtensible Markup Language

Appendix 3: Sources

Interviewees: Alex Cobham, Tax Justice Network; Andres Knobel, Tax Justice Network; Carl Dolan, Transparency Internationa; Chris Taggart, Open Corporates; Jiri Skuhrovec, DataLab.cz / DigiWhist project; Joseph Stead, Christian Aid; JP Rangaswami, Chief Data Officer, Deutche Bank; Robert Palmer, Global Witness.

Participant observation at: Open Government Partnership Global Summit (October 2015, Mexico City), including Open Data Charter workshops; FTC Annual Conference (October 2015, Jakarta).