PARTNERSHIP FOR Open Data

technical report

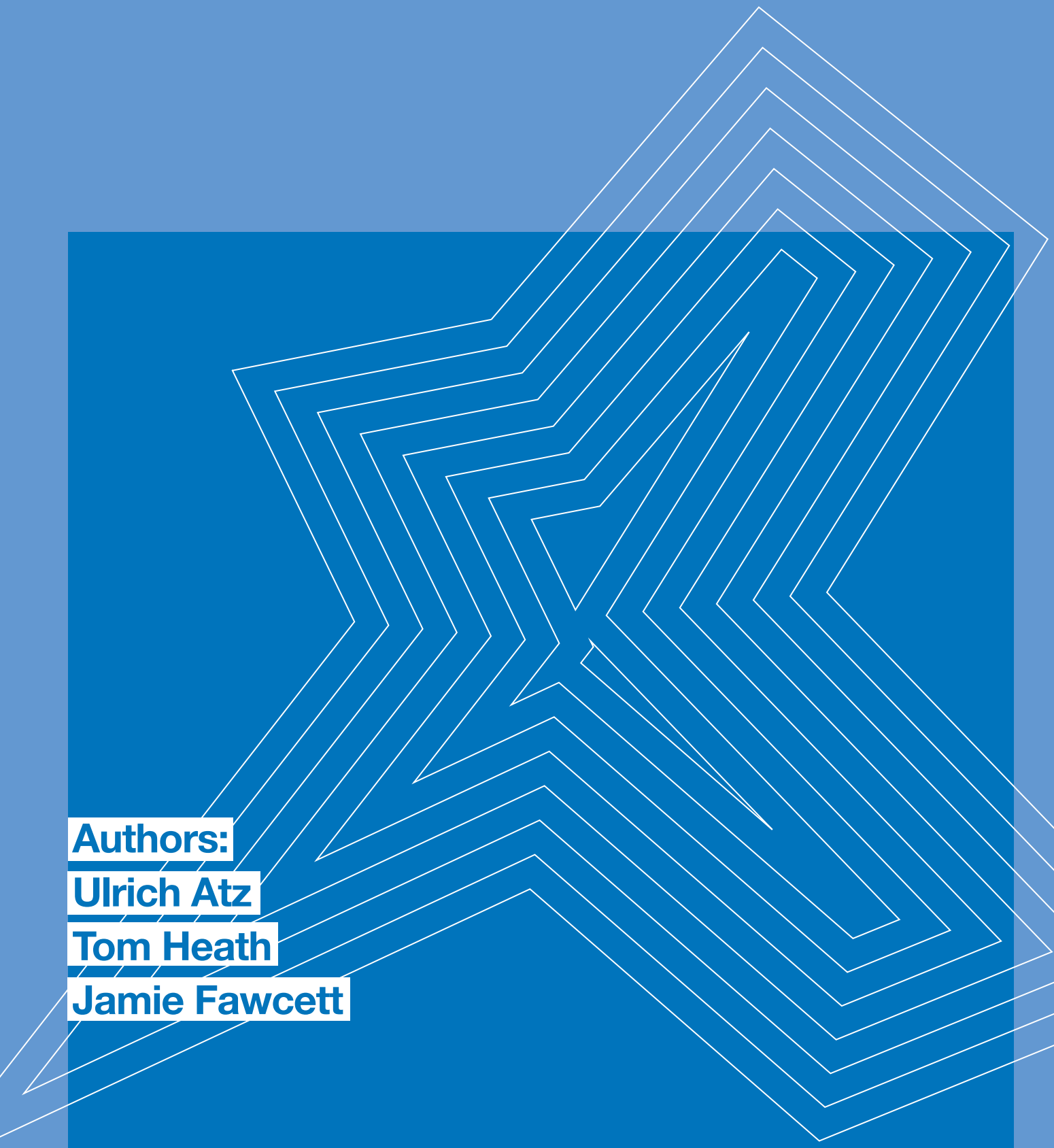
Benchmarking open data automatically

ODI-TR-2015-000

UK20150105

Open Data Institute

Authors:
Ulrich Atz
Tom Heath
Jamie Fawcett



Executive summary

As open data becomes more widespread and useful, so does the need for effective ways to analyse it.

Benchmarking open data means evaluating and ranking countries, organisations and projects, based on how well they use open data in different ways. The process can improve accountability and emphasise best practices among open data projects. It also allows us to understand and communicate how best to use open data for solving problems. Future research and benchmarking exercises will need to happen on a larger scale, at higher frequency and less cost to match the rising demands for evidence.

This paper explores individual dimensions of open data research, and assesses how feasible it would be to conduct automated assessments of them. The four dimensions examined are: open data's context/environment, data, use, and impact. They are taken from the Common Assessment Methods for Open Data (CAF),[1] a standardised methodology for rigorous open data analysis. The paper proposes a comprehensive set of ideal constructs and metrics that could be measured for benchmarking open data: from the existence of laws and licensing as a measure of context, to access to education as a measure of impact.

Recognising that not all of these suggestions are feasible, the paper goes on to make practical recommendations for researchers, developers and policy-makers about how to put automated assessment of open data into practice:

1. Introduce automated assessments of open data quality, e.g. on timeliness, where data and metadata are available.
2. Integrate the automated use of global performance indicators, e.g. internet freedoms, to understand open data's context and environment.
3. When planning open data projects, consider how their design may allow for automated assessments from the outset.

Improving automatic assessment methods for open data may increase its quality and reach, and therefore help to enhance its social, environmental and economic value around the world. For example, putting an emphasis on metadata may ensure that data publishers spend enough time on preparing the data before their release. This paper will help organisations apply benchmarking methods at larger scale, with lower cost and higher frequency.

This paper is part of a series produced by the Open Data Institute, as part of the Partnership for Open Data (POD), funded by the World Bank.

What is open data?

Open data is data that is made available by governments, businesses and individuals for anyone to access, use and share.

What is the Open Data Institute?

The Open Data Institute (ODI) is an independent, non-profit and non-partisan company based in London, UK. The ODI convenes world-class experts from industry, government and academia to collaborate, incubate, nurture and explore new ideas to promote innovation with open data. It was founded by Sir Tim Berners-Lee and Professor Sir Nigel Shadbolt and offers training, membership, research and strategic advice for organisations looking to explore the possibilities of open data.

In its first two years, the ODI has helped to unlock over US$55m in value through the application of open data. With 24 nodes around the world, the ODI has trained more than 500 people from over 25 countries. In 2014, the ODI trained officials from countries including Botswana, Burkina Faso, Chile, Malaysia, Mexico, Moldova, Kyrgyzstan and the UK on the publication and use of open data.

What is the Partnership for Open Data?

The Open Data Institute has joined Open Knowledge and the World Bank in the Partnership for Open Data (POD), a programme designed to help policy-makers and citizens in developing countries to understand and exploit the benefits of open data. The partnership aims to: support developing countries to plan, execute and run open data initiatives; increase reuse of open data in developing countries; and grow the base of evidence on the impact of open data for development. The initial funding comes from The World Bank's Development Grant Facility (WB DGF). Under POD, the ODI has carried out open data readiness assessments, strategic advice, training and technical assistance for low- and middle-income countries across four continents. In 2015, POD will merge with the Open Data for Development (OD4D) network. As part of this new, larger network, the ODI will continue to take a lead in supporting the world's government leaders in implementing open data, and in doing so will continue to publish practical guides and learning materials, such as this series of reports.

Table of contents

1. Introduction to benchmarking open data

There is a global shift towards governments and organisations publishing more open data – that is, data made available for anyone to access, use and share. For example, datacatalogs.org, a meta-list of data portals, counts 390 catalogues across the world.[2] The Open Government Partnership has grown from eight participating countries to 65.[3] In its latest iteration, the Open Data Barometer, a regular survey of government open data readiness, implementation and impact run by the Web Foundation and the Open Data Institute in 2013, targets more than 80 countries.[4]

Policy-makers, civil society groups and businesses demand quantitative evidence for the promised benefits of open data. Many benchmarking efforts are trying to meet these demands. Benchmarking open data means evaluating and ranking countries, organisations and projects based on how well they use open data. The process of benchmarking can improve accountability and emphasise best practices among existing open data projects. Table 1 lists several examples of leading open data benchmarking studies.

Table 1. Examples of open data benchmarking studies

Study	Organisation	Description
E-Gov Survey/ Index [5]	United Nations Public Administration Network	UN E-Gov Survey analyses e-government and e-participation in member states including looking at the publishing of open government data and open data initiatives.
Global Open Data index [6]	Open Data Census/Open Knowledge	Open Data Census explores the openness of a specific set of key government datasets for countries around the world and its Global Open Data Index provides an annual global score comparison between them.
Open Data Barometer [7]	Web Foundation & Open Data Institute	Open Data Barometer measures the distribution and impact of open government data policies and practices around the world, using multidimensional analysis to score countries' overall progress in realising the potential benefits of open data.
Open Data Monitor [8]	European Union Consortium (inc. Open Data Institute	Open Data Monitor assesses trends in the data being published openly by national and regional governments in Europe, through automated analysis of metadata in data catalogues.

Isolated research efforts, however, may lead to duplication, reduce comparability and stifle innovative research. Even case studies that are, by design, unique, benefit from using an overarching framework that embeds their results into the wider context of open data research.

The growing importance of open data means that future research and benchmarking exercises will need to happen on a larger scale, with higher frequency and less cost. Only a quantitative and scalable solution can meet these requirements while factoring in subjective indicators and case study research. This study explores the feasibility of conducting automated assessment of open data, based on the Common Assessment Framework.

2. Adopting the Common Assessment Framework for open data

The Common Assessment Framework (CAF) provides a standardised methodology for a rigorous analysis of the supply, use and impact of open data. The first draft of the framework was developed by the World Wide Web Foundation, the Governance Lab at NYU, the ODI, and other organisations in a workshop held in June, 2014. It aims to loosely coordinate the efforts of researchers and organisations in designing comparable and complementary research.[9] The CAF builds on many of the existing open data benchmarking tools and processes.

The full framework, available in the appendix, consists of four conceptual dimensions:

1. Context/Environment: the context within which open data is being provided. This might be national, in the case of central government's open data, or more specific, in a particular sector such as health, education or transport.

2. Data: the nature and quality of open datasets, i.e. their legal, technical and social openness, relevance and quality. The framework also looks to identify core categories of data that might be evaluated in assessments.

3. Use: the types of users accessing data, the purposes for which the data is used and the activities being undertaken to use it.

4. Impact: the benefits gained from using specific open datasets, or from open data initiatives in general. Benefits can be studied according to social, environmental, political/governance, and economic/commercial dimensions.

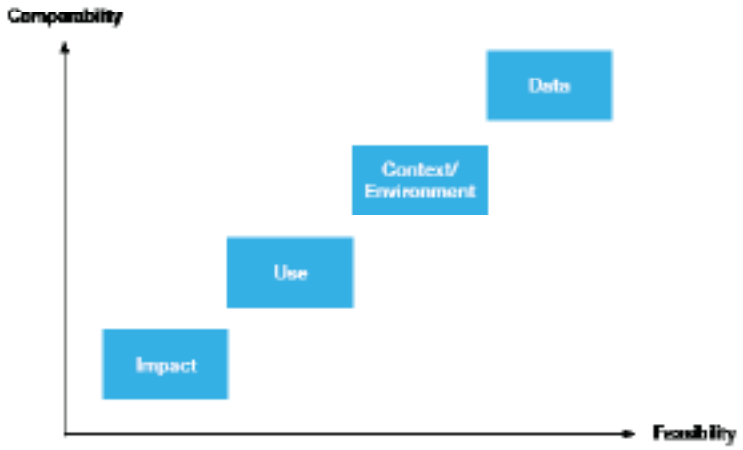
Within each of these dimensions are a number of subcomponents. For example, 'impact'

is split by social, environment, political/governance and economic/commercial categories. Subcomponents are, furthermore, expanded by core questions which aim to direct researchers toward specific aspects to be addressed. For instance, within the 'social' subcomponent of 'impact', comes the question "How can open data be used to increase equality, target resources to citizens, and improve public services?" The framework also lists both examples of potential indicators and existing benchmarking projects.

3. How feasible are automated metrics for the Common Assessment Framework?

The four high level dimensions of the CAF vary widely in their potential for automation. Figure 3.1 provides a conceptual overview of which dimensions are easy to quantify, given the availability of data. They are presented as a hierarchy, based on their potential for automation, but this may simplify the implementation for some scenarios.

Figure 3.1. Feasibility and comparability of the four dimensions, under ideal circumstances



Feasibility *refers to potential application of an automated assessment given ideal availability of data, metadata or corpus such as up-to-date available legislative records or high update frequency of a dataset in a machine-readable format.*

Comparability *refers to the idiosyncratic nature of the dimension, namely how readily the automated assessment may be generalised across other countries, times or domains. User statistics for Transport for London's open data, for example, may be applicable to other large urban agglomerations but limited in other respects. Licences for datasets, especially based on Creative Commons, ought to be globally comparable.*

Table 3.1 provides a brief introduction to each of the dimensions, an overview of the current approaches in each and their potential for automation. This concise analysis allows us to moderate our expectations of the potential for automation in each of the dimensions.

Context/Environment *The context within which open data is being provided. This might be the national context in the case of central Open Government Data, or might be the context in a particular sector such as health, education or transport.*	**When publishers release digital information to a high standard, automated assessments of it are more likely to work well. Careful consideration should go into validating how meaningful the chosen metrics and applicable metrics are for open data.** *Current approaches:* Existing benchmarking organisations provide a range of different measures around the context/ environment of open data. The majority of these are qualitative statements collected through surveys and interviews. However, a few do draw upon quantitative metrics associated with some global performance indicators. *Potential for automation:* There is plenty of scope to develop solid quantitative metrics, especially those based or derived from global performance indices and national government indicators. Automation is highly feasible for a given technical level, for example legislation published on the web. Some methodological questions persist, such as determining the causal impact of open data beyond mere correlations.
Data *The nature and qualities of open datasets. Including the legal, technical and social openness of data, and issues of data relevance and quality. The framework also looks to identify core categories of data that might be evaluated in assessments.*	**Automated analyses of datasets themselves are already a developed aspect of open data benchmarking, but they depend on high-quality metadata.** *Current approaches:* Quantitative metrics such as download statistics are in theory available to open data benchmarking organisations, but are not necessarily consistent or implemented. Automated assessment implementations are being researched through projects like OpenDataMonitor.[10] *Potential for automation:* Given the quantitative nature of data portals and metadata, data benchmarking is the dimension with the highest potential. It is, however, subject to the existence of high-quality metadata in a consistent, standardised and complete format. An additional requirement may be that datasets are organised in international, national or local data portals.

Use *How is data being used and with what possible outcomes? The framework looks at the category of users accessing data, the purposes for which the data will be used, and the activities being undertaken. This part of the framework addresses the 'who, what and why' of open data in use.*	**Assessing how open data is used is feasible for specific cases or applications, but assessing less straightforward use, like secondary reuse of data, poses many challenges.** *Current approaches:* Use of open data is, at least in the first instance, quantifiable through the collection of access statistics of applications, portals and datasets. Many benchmarking organisations actively track the details of use through surveys, interviews and/or case studies. *Potential for automation:* Automation is highly applicable to the primary use of data, subject again to metadata and implementation. The scope for the automated assessment of the purpose of use and reuse throughout the open data ecosystem, however, may be limited. Well-designed systems may be able to quantify uptake ('who') and outcomes ('what') in certain domains. 'Why' people use open data is difficult to observe through behaviours and therefore may not be measurable through automated assessments.
Impact *The benefits gained from using specific open datasets, or from open data initiatives in general. Benefits can be studied according to social, environmental, political/governance, and economic/commercial dimensions.*	**Assessing the impact of open data with automated metrics is difficult, both conceptually and practically. Justifying the causal link between open data and its impact is, while not impossible, a challenging methodological task.** *Current approaches:* To the best of our knowledge, there are hardly any automated metrics that measure the wider economic, social or environmental impacts of open data. Some benchmarking organisations, e.g. the Open Data Barometer, attempt to quantify impact through proxy measures, yet they are typically a comprehensive and costly study. If anywhere, the most promising candidates for measuring impact through automated metrics are found in highly specific use cases. *Potential for automation:* Economic, and to a lesser extent social, political/governance and environmental, impact are in principle quantifiable with an ideal provision of open data. In practice, any automated metrics face the question of how much change can be attributed to open data initiatives. The key here is to establish a credible link between metrics and putative impact.

3.2 Barriers to introducing automated metrics

Beyond the specific limitations set out above, there are universal barriers to introducing automated metrics. These barriers, laid out in Table 3.2, apply to many scenarios because they represent more general issues that people experience when working with data. They provide a conceptual overview that should inform the scope and potential of any specific open data project.

Table 3.2. Barriers to automated metrics

Availability	*Does the data exist?* Quantitative methods rely on the existence of relevant and valid data. The most basic barrier of automated assessment is the lack of data. For example, if no download statistics are available, use is hard to track retrospectively. *Recommendation* Researchers should consider automated methods during early stages of project design. For example, implement tools or site analytics that capture usage data.
Data quality	*Is the data good enough?* Data quality spans a range of issues. It may refer to machine-readable properties, completeness, timeliness, and so forth. For example, assessing how up-to-date open data is requires the metadata to include dates and update frequencies that are standardised. *Recommendation* Researchers, developers and policy-makers should adhere to common data standards as much as possible. For example, data publishers may refer to the Open Data Certificate.
Validity of quantitative metrics	*Is the data meaningful?* Numbers on a dashboard may not necessarily reflect its intended purpose. It is crucial to keep in mind that quantitative metrics are never neutral and carry the implicit decisions by the researcher. For example, tracking the number of datasets in a national catalogue may tell us about the maturity of the country, but often is not a useful proxy for the completeness of open data because even a large amount may miss strategic datasets. *Recommendation* Choosing meaningful metrics requires thinking of the context in which they appear. Researchers should be open to a pragmatic approach, but remain critical of it and carry out revisions if necessary.

4. The ideal approach for benchmarking open data

This section proposes a comprehensive, yet non-exhaustive, set of idealised constructs and metrics.[11] While some of them might not be practical, they are intended to help guide future benchmarking efforts. They do not sketch out an ideal automated benchmarking or policy tool. This could be achieved by weighting and aggregating them into an index, but is beyond the scope of this work. Each of the dimensions is represented by a section (Sections 4.1-4.4) where a table (Tables 4.1-4.4) lists the proposed constructs for each subcomponent with a number of illustrative metrics.

4.1 Context/Environment: measuring the effect of context and environment on open data

Measuring the effect of context and environment on open data requires a broad examination of the legal, technical and organisational context and the environment in which open data is used.

Table 4.1. List of proposed constructs to assess data context and environment

CAF subcomponent	Constructs and idealised metrics
Legal and regulatory	**Open data licensing provision** • *Existence of open licensing framework and policy* • *Textual analysis of licences* • *List of compatible licences* **Functioning right-to-information (RTI) framework** • *Existence of RTI laws* • *Ratio of requests made to information granted* • *Mean time taken for request to be granted* **Functioning public sector information (PSI) reuse policy** • *Existence of law and policy on PSI reuse* • *Statistics on the ease of reuse* • *Extent of adoption of open data legal and regulatory standards* **Internet freedoms, privacy and restriction laws** • *Existence of internet privacy/restriction law and policy* • *Textual analysis of privacy/digital communication laws* • *Score of internet freedoms*

CAF subcomponent	Constructs and idealised metrics
Organisational	**Type/structure of organisations involved** • *Full lists of businesses, government bodies and civil society actors using open data* • *Number of city/regional open data initiatives* • *Count of open data startup incubators* • *Existence/count/size of open data portals* • *Number/size of intermediary open data organisations* **Roles of organisations involved** • *Network analysis of open data actors* **Maturity of the existing open data ecosystem** • *Count of open data actors* • *Number of people or platforms reached by open data* • *Number of organisations involved by date they started using open data* **Continuity of open data usage** • *Rate of uptake in the use/publishing of open data per year* • *Measure of continual usage of open data*
Political will/ Leadership	**Commitment to transparency** • *Government transparency index* • *Measure of centrality of openness in policy* **Government data/technology context** • *Measure of the centrality of technology/data to government policy* • *Level of government online service provision* • *Percentage of government documents that are digitised* • *Existence and strength of information management policy* • *Count of government data roles/positions (high level and overall)* **Engagement of government with other actors around open data** • *Existence of information/data consultations* • *Measure of responsiveness of policy to consultation processes* • *Level of engagement between agencies and developers* **Government promotion of open data goals** • *Textual analysis of government communications (speeches/press releases/publications) for key words* • *Count/percentage of government departments/agencies releasing open data* • *Extent/strength of promotion of PSI reuse*

CAF subcomponent	Constructs and idealised metrics

Technical	**Skills and resources**
	• *Number of data/computer science graduates*
	• *Level of data literacy in civil service/government*
	
	Training and education
	• *Number of education courses around data/technical skills*
	• *Number of educational modules mentioning data management or computer science skills*
	• *Textual analysis of school curricula for data training*
	
	Technical infrastructure
	• *Extent of technology uptake, for example:*
	  • *access to internet*
	  • *access to fibre optic*
	  • *number of mobile phone users*
	  • *number of smartphone users*
	
	• *Cost of technology relative to basic goods*
	• *Level of government and private sector investment in data/technology infrastructure*

Social	**Wider social context** • *Media freedom score* • *Media plurality/diversity* • *Analysis of social media surrounding open data* • *Civil liberties/political freedoms* **Engagement of civil society** • *Number/size of civil society/community/grassroots organisations using and/or promoting open data* • *Level of data literacy amongst population* **Will and leadership within civil society** • *Strength/size of academic output in the field of open data, e.g. number of papers/citations* • *Existence of infomediaries* • *Clout of civil society open data champions*

CAF subcomponent	Constructs and idealised metrics
Economic	**Wider economic context** • *Level of investment in technological innovation from government and private sector* • *Percentage contribution of technology industry to GDP* • *Early stage funding for startups* **Capacity and support** • *Demand/supply for data science/technical positions* • *Level of funding for open data initiatives* • *Firm-level technology absorption* • *Count/size of hackathon/hackday events* • *Count/size of open data marketplaces* **Will and leadership within the private sector** • *Count of private sector open data champions* • *Count of private sector data roles/positions (high level and overall)* • *Number of businesses using/seeking/demanding data*

4.2 Data: measuring the nature and quality of open data

In 2007, a group of open government advocates drafted a set of eight principles of open government data (OGD).[12] For practical reasons, not all of these principles may be assessed in an automated fashion. The list in Table 4.2 goes into more detail. More information about technical aspects of the automated assessment of data catalogues can be found in the reports of the OpenDataMonitor project.[13]

Table 4.2. List of proposed constructs for data

CAF subcomponent	Constructs and idealised metrics
Definitions and dimensions	**Primary** relates to the source of the data. What level of aggregation is appropriate, how to define the original source or how to assess the rawness of data are difficult questions beyond automatic metrics. • *Total number of data catalogues (more is not necessarily better, depending on context)* • *Proportion of dataset distributions in each catalogue that are not listed in any other catalogues* **Accessibility** can be automated for many technical aspects. For example, the distribution of data formats or the number of languages in a catalogue are usually easy to measure. Other, perhaps social aspects, are more difficult to quantify. • *Frequency of dataset distributions with previews* • *Frequency of different languages*

CAF subcomponent	Constructs and idealised metrics
	Non-discriminatory: Measuring if data is available to anyone, with no requirement of registration may be trivial if each data catalogue follows a standard policy. If not, it may still be possible to measure the extent to which open data is available without discrimination via other metadata. • *Proportion of datasets available only via an API* • *Proportion of datasets available in a human-readable file format* **Machine-readable:** It is fairly straightforward to assess all individual datasets on the extent to which they are machine-readable. However, many details may require manual input and/or only emerge as problematic in an actual application. For example, metadata may be machine-readable on a basic level but not include a meaningful schema. • *Frequency of dataset distributions that are machine-readable* • *Frequency of error and warnings generated by, for example, CSVlint (http://csvlint.io, for CSV files)* **Non-proprietary:** Measuring the range of data formats is usually feasible in an automated fashion. The openness of different formats has been measured, for example, with Tim Berners-Lee's 5 stars of open data. • *Frequency of catalogues using specific software platforms* • *Frequency of dataset distributions by file format* **Licence-free:** If each dataset includes an appropriate piece of information regarding its licence, and the number of licences is limited, it may be possible to measure the extent data is available with an open licence. • *Frequency of dataset distributions with an explicitly set license* • *Frequency of datasets distributions with an open license*
Classification / Sectors of datasets	**Sectors of datasets** • *Comparison of published datasets in a sector against list of key sector datasets, for example, based on the Global Open Data Index*[14] • *Cluster analysis of datasets released by sector* **High value datasets**[15] Companies • *Company/business register* Crime and justice • *Crime statistics/safety*

CAF subcomponent	Constructs and idealised metrics
	Earth observation • *Meteorology/weather, agriculture, forestry, fishing, and hunting* Education • *List of schools, performance of schools, digital skills* Energy and environment • *Pollution levels and energy consumption* Finance and contracts • *Transaction spend, contracts let, call for tender, future tenders, local budget, national budget (planned and spent)* Geospatial • *Topography, postcodes, national maps, local maps* Global development • *Aid, food security, extractives, land* Government accountability and democracy • *Government contact points, election results, legislation and statutes, salaries (pay scales), hospitality/gifts* Health • *Prescription data, performance data* Science and research *Genome data, research and educational activity, experiment results* Statistics • *National Statistics, Census, infrastructure, wealth, skills* Social mobility and welfare • *Housing, health insurance and unemployment benefits* Transport and infrastructure • *Public transport timetables, access points broadband penetration*

CAF subcomponent	Constructs and idealised metrics
Quality	**Completeness** may be measured automatically, however, any metric has to be reviewed over time. The set of open data evolves as more is understood about its impact and usefulness. It may be possible to compare completeness against a pre-defined universe of open data (see above). • *Frequency of catalogued datasets* • *Size of datasets and catalogues* • *Frequency of catalogues by sector of publishing organisation* **Timeliness:** Up-to-date catalogues and timely data can be measured automatically, provided the metadata is standardised. • *Median days since latest dataset update* • *Proportion of datasets with stated update frequency* **Metadata:** i.e. data completeness, standardisation and relevance. • *Adherence to a standard such as the Dublin Core Metadata Initiative (DCMI)* • *Proportion of data file links that are broken* • *Number of fields in the metadata record that are populated* • *Open Data Certificate level of the dataset*[16]

4.3 Use: measuring how and why open data is being used

Measuring how open data is used requires an examination of:

- who the users of open data are,
- what data they are using,
- why they are using it, and
- how they are using it to inform their projects.

Table 4.3. List of proposed constructs for use

CAF subcomponent	Constructs and idealised metrics
Users	**Current users** • *Number of users/download statistics for each catalogue* • *Number of users/download statistics for each dataset* • *Analysis of user demographics/sectors*

CAF subcomponent	Constructs and idealised metrics
	Potential users • *Profile of existing users across demographics/sectors* • *Number of users using closed government data* • *Measure of size/scope of proprietary data usage* • *Value/size/scope of proprietary data market* **Non-users** • *Affordability of data services/infrastructure for various sized of businesses* • *Number of actors who have stopped releasing/using open data*
Purpose	**Perceived motives** • *Percentage using open data in current field versus percentage trying to enter a new field* • *Observed behaviour: increased value, lowered cost, improved experience, disrupted or enhanced existing activities* • *Type of project: business/social/environmental* **Ambition and goals** • *Scale of outputs: local, national, international* • *Percentage of those who publish/report results* • *Percentage of revenue types (premium, freemium etc)*
Activities	**Uses/outputs** • *Count/size of secondary open data* • *Analysis of applications and related tools* • *Type of project outputs: report, data, software etc.* **Sectors** • *Sector/type of datasets most published* • *Sector/type of datasets most used* • *Sector/type of actors most involved* • *Sector/type of outputs most produced (apps, reports, etc)*

4.4 Impact: measuring the benefits of open data

Measuring the impact of open data is perhaps the most important and most difficult task in benchmarking open data. Demonstrating social, environmental, political and economic impact in specific settings is of most use if it is possible to show how the impact may be generalised. Demonstrating impact for a wider scope depends on establishing a credible causal link between the open data initiative and its putative impact. The list of challenges that open data may support spans all areas, hence the list of proposed constructs below remains high-level.

Table 4.4. List of proposed constructs for impact

CAF subcomponent	High-level constructs
Social	**Education** • *Access to education* • *Quality of education* • *Lifelong learning and development opportunities* **Health** • *Combating disease and increasing life expectancy* • *Promotion of healthy lives and well-being* • *Development of the healthcare system and healthcare delivery* **Human settlements** • *Sustainable land use, building and infrastructure planning* • *Ability to house citizens* • *Ability to manage urbanisation* **Transportation** • *Access to transportation* • *Increased efficiency of transportation* • *Transport infrastructure* **Social development** • *Gender equality and empowerment of women* • *Protection of vulnerable society members* • *Social inequality* • *Personal financial management* • *Social and economic security*
Environmental	**Environment and natural resources management** • *Preservation of the environment and habitats* • *Resilience to natural disasters and climate change* • *Sustainability* • *Pollution* **Food and water** • *Access to affordable and healthy food* • *Access to clean water* • *Sustainable agriculture* **Sanitation and waste management** • *Access to proper sanitation* • *Waste management capability* • *Recycling* **Energy** • *Renewable energy* • *Efficiency in the delivery of energy* • *Reliability of energy in homes*

CAF subcomponent	High-level constructs
Political/ Governance	**Governmental efficiency** • *Public services* • *Reduced crime and violence* **Governmental accountability** • *Reduced government corruption* • *Attitudinal changes toward government agencies* **Civic engagement** • *Political freedom* • *Political participation*
Economic/ Commercial	**Economic prosperity** • *Innovation and entrepreneurship* • *Wealth and inequality* • *Employment and unemployment statistics* • *Job creation* • *Trade and investment* **Growth in the open data landscape** • *Total number of open data businesses* • *Size/profit of open data businesses* • *Number of new jobs created in the (open) data sector* • *Size of tax revenue generated from open data companies*

5. Towards a pragmatic, automated approach to benchmarking open data

The automation of many metrics listed above is currently not feasible. This is in part due to the barriers to automation discussed in this study. It is unlikely that in the foreseeable future there will be reasonable proxy measures for some constructs. There are also many broader practical limitations, for example, incomplete or substandard metadata, that are common in many datasets. It is therefore important to manage expectations surrounding what is possible with regards to automation.

In order to operationalise these metrics, it is necessary to identify sources of data, and, so far, they fall primarily into three categories:

1. Global Performance Indices (GPI): GPI's are useful sources of data for automated metrics, given that they are often comprehensive in country coverage, published online, reliable and available on a wide range of topics (at least 150 indices exist).[17] Examples of other sources include the World Bank Data,[18] UNdata[19] and OECD data[20] platforms.

2.Government data: In many cases metrics rely on (open) government data for a wide range of information regarding its own makeup, practices and legislation. UK examples of sources for such data include legislation.gov,[21] government announcements[22] and data.gov.uk.[23]

3. Portal metadata: Portal metadata is essential for analysis of the data dimension of the CAF. Portals might be local, regional, national or international in scale with appropriate granularity or aggregation. Portals for France, for example, include the City of Paris open data,[24] Région Île-de-France open data,[25] data.gouv.fr[26] and EU open data.[27]

Note: For sources 2 and 3, government data and portal metadata, there are a number of caveats to automation:

a. Tools will need to be pointed toward the relevant sources by researchers requiring an initial investment in resources.

b. In general, automation assumes a collaboration between the data providers and excludes other forms of collection such as scraping.

c. To be useful, the data must be relevant and of sufficient quality.

The next section demonstrates how new or existing benchmarking organisations can create automated assessment methods measuring metrics within CAF's four dimensions. These metrics should be able to supplement and streamline existing processes in a viable and useful way.

5.1 Measuring context/environment: the scope for automation

To measure the context and environment of open data, we can often rely on existing Global Performance Indices. GPIs are in many cases available for all, or nearly all, countries and produced yearly, which supports their automated integration. Table 5.1 lists a few examples used in the Open Data Barometer.

Table 5.1. Examples of existing metrics using GPIs mapped to CAF constructs

Construct	Example metric	Source
Government data/ technology context	Importance of ICT to government vision (Variable 8.01)	World Economic Forum global information technology reports [28]
Technical infrastructure	Internet users per 100 people (IT.NET.USER.P2[29])	World Bank Data[30]
Wider social context	Civil liberties rating	Freedom House Political Freedoms and Civil Liberties Index[31]
Capacity and support	Firm-level technology absorption (Variable 9.02)	World Economic Forum Global Competitiveness Index[32]

Table 5.2 shows examples of data sources that are based on government open data portals.

Table 5.2. Examples of potential sources for CAF constructs for different countries

Construct(s)	Metric(s)	Example countries	Sources
Legal and regulatory constructs	Textual analysis of laws	Kenya Sweden	Laws of Kenya database[33] Laws and regulations of Sweden[34]

Construct(s)	Metric(s)	Example countries	Sources
RTI laws	Measure of effectiveness	Brazil USA	Access to information statistics[35] Freedom of information statistics[36]
Government promotion of open data goals	Textual analysis of government communications	Australia South Africa	Government media releases[37] Department of Communications subscriptions[38]

5.2 Measuring data quality: the scope for automation

Pragmatic automated metrics exist for metadata that stems from data portals such as CKAN, Socrata, OpenDataSoft or DataPress. A detailed implementation is the monitoring platform being developed by the OpenDataMonitor project, using analysis and visualisation techniques to give insights into open data deployment across Europe.[39] The platform harvests metadata from local, regional and national open data hubs, and includes an extensive list of automated metrics.[40]

5.3 Measuring data use: the scope for automation

Primary use of open data is fairly easy to quantify if the data is linked to widespread digital analytics tools, and example being the site analytics of the UK data portal data.gov.uk. Metadata from portals ought to provide a simple way to monitor download and user statistics with a high granularity, for example. However, it is much more difficult to automatically assess the use of open data in secondary instances such as reuse of data. In some cases, the open data value chain can be extensive.

5.4 Measuring data impact: the scope for automation

As has been discussed in detail, measuring impact with an automated approach is inherently difficult. Most likely, researchers will have to rely on proxy indicators because high-level constructs such as reduced corruption are hard to quantify. There will be several elements in an open data impact evaluation that require the analytical reasoning of a researcher. In fact,

the literature on impact evaluation is vast and open data initiatives may be able to adapt many of the leading practices.

This is not to say that in some cases an automated assessment is not attainable. However, it is the researcher's or organisation's responsibility to justify why such metrics are a valid representation of the open data impact.

6. Recommendations for benchmarking organisations

Given the varied scope and nature of benchmarking organisations, our recommendations can only be generalised. The lists provided in the previous sections serve as guidelines, with some more concrete suggestions. Based on this analysis and previous work, more automated assessments should be possible in the future. Moreover, automated metrics can offer an opportunity for larger scale, more frequent and less expensive assessments.

The following recommendations are for new and existing benchmarking organisations:

1. **Introduce automated assessments of open data quality, where data and metadata are available**

   The analysis of data's nature and quality has the highest feasibility for automation. Data are typically quantitative, in some form, and are associated with metadata, i.e. data about data. This means that if data is provided in, for example, a hosting solution such as CKAN, Socrata, DataPress or OpenDataSoft, researchers can build automated assessments on top of these standardised platforms. The OpenDataMonitor project offers examples of how this works.

2. **Integrate the automated use of Global Performance Indicators (GPIs)**

   In the last decade, the availability of GPIs has risen dramatically. While many may be unrelated to open data, there are several that may help to understand the context and environment of open data initiatives. The advantages are that these indicators are usually available for free, with regular updates, for many or all countries and based on deliberate methodologies. On its own, a GPI may not be sufficient for a benchmarking approach, but, as part of a wider scope, there is potential for automation.

3. **Adopt an approach that considers the automated assessment of open data early on in their planning**

In many cases, automation fails for the most basic of requirements: the availability of data. Without relevant and valid data sources, there will not be automated methods. It is therefore crucial for researchers, developers and policy-makers to consider automation at the design phase of their projects. Small changes such as the collection of key metadata can make the difference whether an automated assessment is feasible later on. In general, these considerations have wider benefits, for example, putting an emphasis on metadata may ensure that data publishers spend enough time on preparing the data before its release.

We invite researchers to share their approaches to data analysis and automation.[41] As the open data landscape evolves, established methods will improve, proposed methods will become more feasible and new methods will emerge. Research in open data, similar to open data itself, should therefore lead by example and stimulate the network effect of sharing leading practices with the community.

Glossary

Methodology box 1. What is a construct?

The term ‘construct’ in social research is commonly used to denote an underlying theme, concept or subject that cannot be measured directly. For example, a construct was identified for the legal context/environment open data licensing provisions, which may be operationalised by a list of compatible licences or a textual analysis of licences. Simple constructs may be measured with one or a few metrics, while more complex ones such as internet freedoms may require a whole battery of indicators.

The quality of a metric or indicator is reflected in its construct validity. For example, how well does the number of data/computer science higher level graduates reflect the availability of technical skills related to open data? Validity is typically represented as accuracy or the degree to which an indicator measures what it purports it does. It refers to how far inferences can be justified from the chosen indicators. Sometimes it is called a ‘labelling issue’ or how well your operationalisation reflects what you are trying to measure.

You can find more information and background on constructs and validity in:
Alasuutari, P., Bickman, L., & Brannen, J. (Eds.). (2008). The SAGE handbook of social research methods. Sage.

Endnotes

1. The first draft of the framework was developed by the World Wide Web Foundation, the Governance Lab at NYU, the ODI, and other organisations in a workshop held in June 2014. http://opendataresearch.org/sites/default/files/posts/Common%20Assessment%20Workshop%20Report.pdf

2. Data Catalogs, http://datacatalogs.org, accessed on 2014-11-10.

3. Open Government Partnership, http://www.opengovpartnership.org, accessed on 2014-12-18.

4. Open Data Research Network, *Research Project: Open Data Barometer*, http://www.opendataresearch.org/project/2013/odb, accessed on 2014-12-18.

5. United Nations Public Administration Network (2014). *UN e-Government Surveys*. Available at http://www.unpan.org/egovkb/global_reports/08report.htm, accessed on 2014-12-18.

6. Open Knowledge, *Open Data Census*, http://census.okfn.org, accessed on 2014-12-18.

7. Open Data Research Network (2013). *Open Data Barometer*. Available at http://www.opendataresearch.org/barometer, accessed on 2014-12-18.

8. OpenDataMonitor, http://project.opendatamonitor.eu, accessed on 2014-12-18.

9. World Wide Web Foundation & GovLab. (2014). *Towards common methods for assessing open data: workshop report & draft framework*. Available at http://opendataresearch.org/sites/default/files/posts/Common%20Assessment%20Workshop%20Report.pdf, accessed on 2014-12-18.

10. OpenDataMonitor, http://project.opendatamonitor.eu/, accessed on 2014-12-18.

11. ODI, *Open Data Certificate*, https://certificates.theodi.org/, accessed on 2014-12-18.

12. See methodology box 1 in the glossary for more information.

13. Some general indicator examples can be found here: http://www.epsiplatform.eu/content/psi-scoreboard-indicator-list, , accessed on 2014-12-18.

14. E.g. Freedom House, *2013 Global Scores* https://freedomhouse.org/report/freedom-net-2013-global-scores, accessed on 2014-12-18.

15. E.g. Transparency International, *2014 Corruptions Perception Index*, http://www.transparency.org/cpi2014/results, accessed on 2014-12-18.

16. E.g. Reporters Without Borders, *World Press Freedom Index 2014*, http://rsf.org/index2014/en-index2014.php, accessed on 2014-12-18

17. The Annotated 8 Principles of Open Government Data, http://opengovdata.org, accessed on 2014-12-18.

18. OpenDataMonitor, http://project.opendatamonitor.eu/, accessed on 2014-12-18.

19. Open Knowledge, *Open Data Census*, http://census.okfn.org/, accessed on 2014-12-18.

20. Taken from the G8 open data charter, available at https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex, accessed on 2014-12-18.

21. ODI, Open Data Certificate, https://certificates.theodi.org/, accessed on 2014-12-18.

22. A full list of indices is awaiting publication: figure drawn from Kelley, J. G., & Simmons, B. A. (2014). The Power of Performance Indicators: Rankings, Ratings and Reactivity in *International Relations* (SSRN Scholarly Paper No. ID 2451319). Rochester, NY: Social Science Research Network. Retrieved 2014-12-18 from http://papers.ssrn.com/abstract=2451319

23. World Bank, *Data,* http://data.worldbank.org/, accessed on 2014-12-18.

24. UNdata, http://data.un.org, accessed on 2014-12-18.

25. OECD, *Data*, http://data.oecd.org, accessed on 2014-12-18.

26. Legislation.gov.uk, http://www.legislation.gov.uk, accessed on 2014-12-18.

27. Gov.uk, *Announcements*, https://www.gov.uk/government/announcements, accessed on 2014-12-18.

28. Data.go.uk, http://data.gov.uk, accessed on 2014-12-18.

29. ParisData, http://opendata.paris.fr/, accessed on 2014-12-18.

30. Data.Iledefrance.fr, http://data.iledefrance.fr/explore/, accessed on 2014-12-18.

31. Data.gouv.fr, https://www.data.gouv.fr/fr/, accessed on 2014-12-18.

32. European Union Open Data Portal, https://open-data.europa.eu/en/data/, accessed on 2014-12-18.

33. World Economic Forum (2012). *The Global Information Technology Report 2013 Data Platform*. Available at http://www.weforum.org/global-information-technology-report-2013-data-platform, accessed on 2014-12-18.

34. World Bank, Internet users (per 100 people), http://data.worldbank.org/indicator/IT.NET.USER.P2, accessed on 2014-12-18.

35. World Bank, *Data*, http://data.worldbank.org/, accessed on 2014-12-18.

36. Freedom House, *Freedom in the World*, https://freedomhouse.org/report-types/freedom-world, accessed on 2014-12-18.

37. World Economic Forum (2014). *The Global Competitiveness Report 2014-2015.* Available at http://reports.weforum.org/global-competitiveness-report-2014-2015, accessed on 2014-12-18.

33. Kenya Law, *The Laws of Kenya*, http://www.kenyalaw.org:8181/exist/kenyalex/index.xql, accessed on 2014-12-18.

39. Lagrummet.se, http://www.lagrummet.se, accessed on 2014-12-18.

40. Acessoainformacao.gov.br, http://www.acessoainformacao.gov.br, accessed on 2014-12-18.

41. FOIA.gov, http://www.foia.gov/developer.html, accessed on 2014-12-18.

42. Australia.gov.au, *Government Media Releases*, http://www.australia.gov.au/news-and-media/government-media-releases, accessed on 2014-12-18

43. Department of Communications SA, *Subscriptions*, http://www.gcis.gov.za/content/newsroom/subscriptions, accessed on 2014-12-18.

44. OpenDataMonitor, http://project.opendatamonitor.eu, accessed on 2014-12-18.

45. Atz, U., Heath, T., Heil, M., Hardinges, J., & Fawcett, J. (2014) Best practice visualisation, dashboard and key figures report. *OpenDataMonitor*. Open Data Institute, London, UK. Available at http://project.opendatamonitor.eu/wp-content/uploads/deliverable/OpenDataMonitor_611988_D2.3-Best-practice-visualisation,-dashboard-and-key-figures-report.pdf, accessed on 2014-12-18.

46. Data.gov.uk, Site Usage, http://data.gov.uk/data/site-usage#totals, accessed on 2014-12-18.

47. Please contact us at research@theodi.org.

Authors:
Ulrich Atz
Tom Heath
Jamie Fawcett

Editing:
Anna Scott
Kathryn Corrick

Production:
Phil Lang

Design
Adrian Philpott