Shareable by Default

Creating resilient data ecosystems

May 2016

open data institute

the answer company™
THOMSON REUTERS®

Introduction

This is the second white paper produced from the partnership between Thomson Reuters and the Open Data Institute. Through that partnership we are seeking to work together on collaborative projects exploring data techniques that benefit the wider open data community.

In the last paper we looked at best practices in open identifiers, a crucial topic for how data can be effectively published, retrieved, reused and linked. In this paper we seek to go further, looking at the ways in which data must be managed to allow it to break out of the many silos that exist within and between organisations, and even across open data repositories. This will enable anyone or any organisation to truly benefit from the whole data spectrum they are able to utilise — whether that data is open, shared or closed.

If you are a chief data officer, or in any way responsible for how a dataset is managed (either from the top down or the ground up), then our hope is that you will read this paper and come away with a set of frameworks and approaches to apply. We want to strongly make the case that the challenges of making data 'shareable by default' are more likely to be around issues of perspective and culture than technology. If everyone takes a position of collaboration and reuse, no matter where their data sits in the data spectrum, everyone will benefit.

We would like to acknowledge and thank all those within the Open Data Institute and Thomson Reuters who contributed to and reviewed the material herein.

Signed,

Debra Walton
Chief Product and Content Officer
Financial & Risk
Thomson Reuters

Jeni Tennison
Technical Director and Deputy CEO
Open Data Institute

Management summary

Data has more potential value when it can be shared or opened. It can be used by a number of different stakeholders and selected partners, within or outside an organisation, for a variety of applications to gain new analytical insight and to build new products and services.

Open innovation requires data that is *shareable*. Invisible frameworks that govern the structure and management of data must be interoperable to facilitate the value that can be extracted from it. This does not mean all data should become open to everyone. Data must be treated as infrastructure, with special attention given to the features of data and the surrounding mechanisms that make data shareable, regardless of whether that data is closed, shared or open.

Drawing from our experience in building and working with data services and ecosystems, in this paper we examine the features and components necessary for effective data management and use. Whilst data can power new ways of thinking, provide a competitive advantage and drive value creation, these technical, organisational and cultural challenges must be overcome to truly reap the benefits of a data-driven approach.

The shareable data checklist:

- **Communicate and document meaning** to ensure that data can be understood by others
- **Describe data provenance** to determine the context of its origin and fitness for use
- **Describe access and usage rights** to ensure that the right people gain access to the data and that all aspects of privacy are respected
- **Publish a description of the data in a data catalogue** to enable search and discovery
- **Provide efficient ways for users to access data on demand**, using a variety of tools
- **Use common standards strategically** to ensure interoperability and take advantage of network effects
- **Design data for everyone** so that its potential use is not restricted to the current community and application
- **Cultivate an open, collaborative culture** to encourage the creation of data that is shareable by default and the reuse of data

Data has more potential value when it can be shared or opened

Many more organisations are becoming data businesses:

- Do you collect data to streamline your operations?
- Are you using data from your suppliers and customers to manage those relationships?
- Is there an expectation that you are providing data to others, whether that be your customers or your partners?
- Do you use data analytics to gain competitive advantage?

Various platforms and analytical capabilities can sit on top of available data, promising the opportunity for insight. But in order to meet the needs of a constantly changing technological environment, businesses must also become more fluid and adaptable in their use of data. Businesses are now working in a context of increasing transparency, where sharing data internally and opening data externally is unlocking value through open innovation.

All these objectives hinge upon the nature of the data itself. Data that is well described, well managed, easy to find and easy to consume can be shared effectively within an organisation and opened to external partners to gain new insights and build better products and services. **The potential value of data is unlocked when it is shared or openly published, enabling open innovation.**

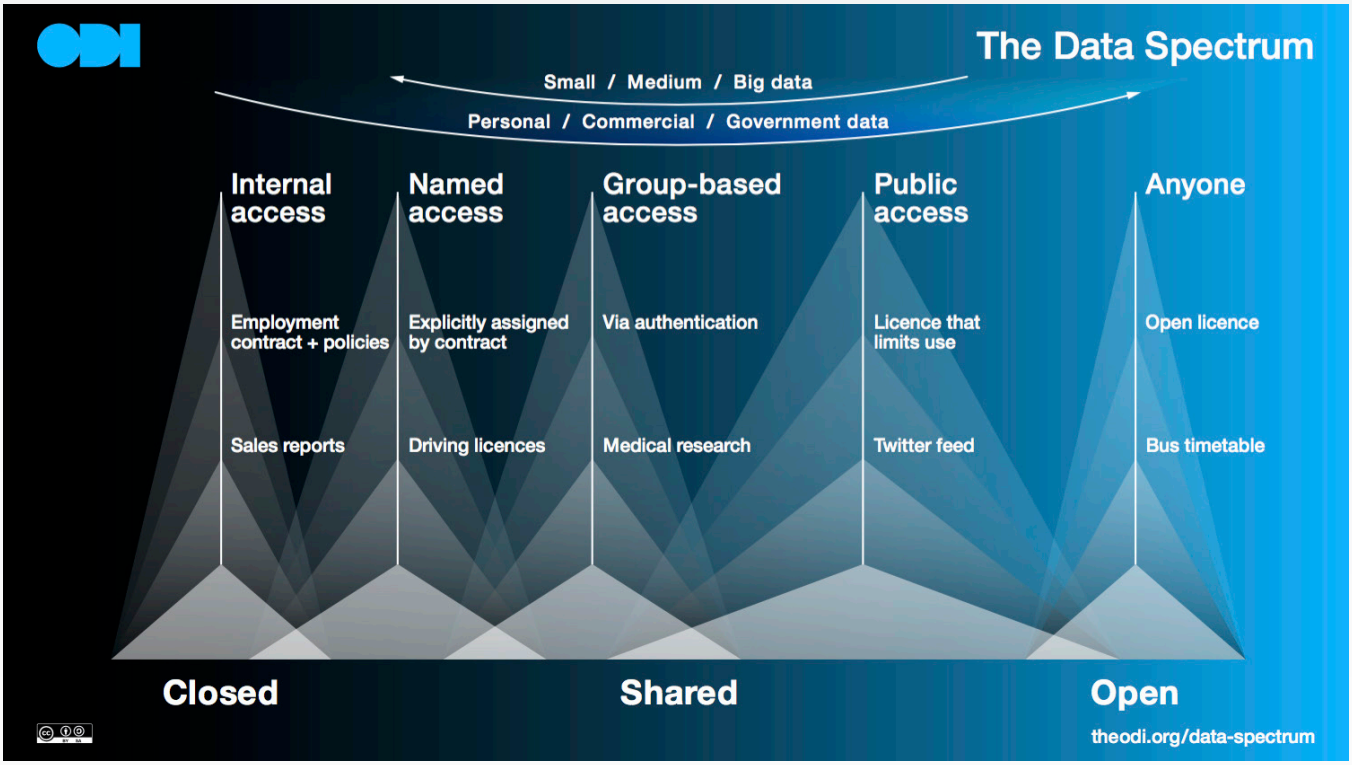
However, **for data to be effectively shared or opened it must first be *shareable***. Above and beyond the relatively simple mechanics of transferring data, this means thinking about the context in which data is made available and ensuring that the surrounding mechanisms enable its reuse. **Prior work to make data shareable lowers the cost of reuse** for consumers. Data publishers and consumers will need to minimise the expected cost of this work to maximise its benefits.

The characteristics that make data shareable are constant regardless of the type of data involved (they can apply to a document, list of numbers, algorithm or mathematical function) or whether that data is structured or not. For example, an item of unstructured data must have access rights and in most cases some temporal context such as the date of creation, just as structured data must.

Shareable data is prepared in advance so as to be readily reusable in diverse contexts. The more shareable data is, the more potential value it holds, because it is easier to combine across the data spectrum, from closed, through shared, to open. This provides organisations with more flexibility in how they get value from data and broadens the opportunity for everyone to benefit by making it easier to increase data openness.

Data that is well described, well managed, easy to find and easy to consume can be shared effectively within an organisation and opened to external partners to gain new insights and build better products and services. The potential value of data is unlocked when it is shared or openly published, enabling open innovation.

WHAT ARE OPEN, CLOSED AND SHARED DATA?



Open data is **data that anyone can access, use and share** — commercially and non-commercially. For data to be considered open, it must be **accessible**, which usually means published on the web, and be available in a **machine-readable** format. Such data should be accompanied by an open data licence.

Some datasets contain sensitive information, which means it may only be shared with a smaller audience, or kept closed. Data might be limited to:

- **Internal access**[1] — closed data, limited to access by its subject, owner or holder
- **Named access** — shared with specific people or organisations for a defined purpose
- **Group-based access** — shared with special groups or people who meet certain criteria, e.g., researchers
- **Public access** — shared with everyone, but under terms and conditions that are not open

Data can move in both directions along the spectrum, and new data assets can be produced that reside in different parts of the spectrum. Closed health data may be anonymised and released as open data, while open data can be mixed with other internal sources, adding context that allows data to be reused in new ways, but only by licenced partners.

Shareable by default

Open innovation, both within the boundaries of an organisation and externally, requires data that is managed to be shareable *before* it is shared or opened. Many organisations are appointing chief data officers (CDOs) to oversee their strategic use of data, including how it can be used to drive innovation. But to do this, CDOs are faced with a number of challenges:

- **Building the organisational capacity** to use data in more creative and innovative ways, by acquiring and nurturing talent, breaking data silos and using platforms designed to help avert the solidification of data assets into solution-specific forms
- **Understanding the variety of datasets** that are either in use or available for use by the organisation
- **Improving data governance** so that data is managed as an asset
- **Ensuring regulatory compliance** within an evolving regulatory landscape
- **Protecting data privacy** by managing complex internal and external access requirements

By ensuring that organisational data is *shareable by default,* CDOs will be able to address these challenges systematically. Shareable data enables the creation of a data ecosystem that can support data reuse by a wider variety of internal data consumers, business partners and technology providers. It also helps prepare the organisation for potential collaborations with external participants and for dealing with, and making better use of, a 'mixed economy' of closed, shared and open data. Making data shareable further requires data publishers to be vigilant in regard to regulatory compliance issues and management of access rights.

[1] These and other data-related terms are described in more detail in the ODI Data Lexicon

Treating data as infrastructure

Physical infrastructures such as roads, bridges and rail are built to support transportation mechanisms. But as they form connections, they also support trade, social exchanges and a whole range of other activities. Data similarly is a form of infrastructure[2, 3]. Data could be of various types and formats, originate from a number of different authorities, and be distributed in different physical locations and stored in different technological environments. However, individual datasets can be linked to form connections across various sectors and organisations. Processes and capabilities can be built on top of the data and the connections between them, supporting a wide range of applications. The data itself thus forms a network, a layer that sits on top of computing technology, supporting a variety of functions. Producing shareable data, which can be connected and exchanged, fosters the creation of this kind of infrastructure.

What makes data shareable?

While physical IT infrastructure and tools are often prominent in the discussion on how best to obtain value from data, conversations around the logical frameworks needed to successfully share, understand and translate data are often overlooked. Take the ubiquitous cardboard box used to ship mixed goods by post or by courier. The efficiency of the box is crucially dependent on the barcode and address labels stuck onto it that link the box to data guiding its passage. The ability of the various systems involved to share, understand and use the data corresponding to those labels is equally crucial.

Equivalent logical frameworks govern how we structure and manage our data 'boxes' and determine the value that can be extracted from them. For data to be shareable we have to design and adopt this framework. Many of its features are driven by metadata: 'data about data' that helps communicate the context in which that data was created and can be used to describe the characteristics of data in a human and machine-readable way. Metadata is therefore central to the structural design of the logical framework. It essentially determines what can be known about the delivery box without opening it. Metadata enables search, discovery and passage of data.

The decision on what metadata should include — the context that needs to be communicated — will be guided by the needs of data users and the information model used: an organised, stable and shareable representation of the concepts, relationships, constraints, rules and operations that govern the data involved. Establishing an information model helps to maintain conceptual consistency for a dataset.

What do we mean by metadata?

Take an example data element, the number '40.24'. On its own it is meaningless: To be useful the full context must also be available (e.g., top of book price in USD for TRI common stock on NYSE at 15:43:00 on 11 November 2015 provided under a free 15-minute-delay policy, as published by Bloomberg at 15:58:10 on 11 November 2015). Providing as much detail as possible about its context in the form of metadata to that data element ensures that whatever a recipient needs to do, they are able to independently judge the appropriateness of the data to their purpose.

1. COMMUNICATE AND DOCUMENT MEANING

In order to effectively use data, especially in a collaborative environment such as the web, it is vital to comprehend exactly what that data is about and how it originated, and is one of the main challenges in sharing data.

For example, like the barcode on the cardboard box, identifiers are labels that refer to the unique object being discussed — such as a product, person or company or even a dataset or document. Successfully expressing and interpreting identities in data is the foundation for linked data and key to obtaining value from data for both publishers and consumers. This requires the use of resolvable, stable and well-documented identifiers[4]. Thomson Reuters Open PermIDs are open, permanent and machine-readable identifiers where underlying attributes capture the context of the identity they each represent. Employing such identifiers to communicate meaning in data ensures the user's ability to understand whether the data suits their purpose.

Recommendations:

- Publish human and machine-readable schemas that describe the structure and content of the dataset
- Describe the codes and meaning of data items, e.g., by publishing machine-readable data about them
- Use resolvable, stable, open identifiers within the dataset to help describe the entities that the data is describing

[2] Data-driven Innovation for Growth and Well-being — OECD

[3] We need to strengthen our data infrastructure

4 The importance of identifiers is discussed in our previous white paper – Creating Value with Identifiers in an Open Data World

2. DESCRIBE DATA PROVENANCE

Provenance describes the means by which data has been collected, processed and managed. Understanding the provenance of a dataset can help users understand whether it has sufficient quality or authority for their purposes. For example, data about countries that is crowd-sourced through Wikipedia might be sufficient for one analysis, whereas a more official register, e.g., from the UN or a national authority, might be preferred for other applications. Obtaining conflicting data from two sources ultimately necessitates the selection of the authority more suited to the application. News organisations may use a different list of independent states altogether to reflect emerging and newly established ones. Awareness of the authorities or activities involved in creating, updating, influencing or delivering data can significantly affect its value. By detailing provenance, publishers help users overcome the challenge of determining the quality of the data — its fitness for the use to which it is being put. This may depend on a number of factors including completeness, accuracy, timeliness, consistency, formats and indeed authorities and provenance.

Recommendations:

- Document the processes by which data has been collected and governed to help reusers understand how it has been curated
- Provide details about any quality control processes that have been applied to the data, to help describe its accuracy
- Publish documentation that describes how often the data will be updated and the processes by which it might be revised
- Provide machine-readable metadata about the provenance of the dataset

3. DESCRIBE ACCESS AND USAGE RIGHTS

As we place our data on different points of the spectrum ranging from closed to open, our logical framework must implement what that signifies. Users may have different rights to view, edit and use data, largely determined by where they lie in relation to organisational boundaries. Users will have alternative views of the data landscape, according to the content that they have access to, or that they can see exists. In an environment where a variety of data types and sources are employed, and where users have varying degrees of access at different points of a dataset lifecycle, communicating these rights and complying with obligations can be difficult.

In making data shareable, data publishers must also take the necessary steps to identify and translate rights information for data that is either commissioned or derived, and also include rights information for any derivations of the data currently being published. Making these rights machine-readable (using emerging standards such as ODRL) will make these processes much more efficient.

Recommendations

- Ensure that shared data is published under an appropriate data usage policy or agreement that governs how it can be accessed, used and shared
- Ensure that data that is openly published is done so under a standard open licence
- Use standards such as ODRL to provide machine-readable descriptions of access rights

Measuring the effectiveness of shareable data

For open data, an open data certificate measures how effectively a dataset is published for ease of reuse. A certificate reviews more than just technical issues, recognising the importance of rights and licencing, documentation and guarantees around availability. The context provided about the dataset is used to award a rating that reflects the effort invested in publishing the data.

Certificates can be awarded in a variety of ways including self-assessment and through a formal audit by the ODI. The certification process distills a variety of best practices relating to data publication on the web to support widespread reuse. Organisations may wish to consider similarly certifying their own shared or closed data, perhaps with the addition of relevant criteria, to help provide a ready reference for internal data consumers.

4. PUBLISH A DESCRIPTION OF THE DATA IN A DATA CATALOGUE

For data to be used it must be discoverable. A data catalogue supports discovery by using dataset metadata to allow potential users to search and browse for reusable data, and can define the nature of any interaction between data publishers and users

An organisation-wide data catalogue can become a strategic resource that helps with the discovery and governance of data. The catalogue may also include external datasets that are being consumed by the organisation.

The descriptions of items in the catalogue should help reusers identify if they can meet their needs, e.g., by providing an indication of quality, provenance, usage rights and pointers to other ways in which the data has been used. In many circumstances, successful use of data relies on a well-designed data catalogue. While metadata might include information on rights, separate mechanisms must control and determine what content a user has access to. This may also form part of the data catalogue function. For data publishers, catalogues can therefore power more than discoverability.

Recommendations

- Create a data catalogue that captures metadata about the datasets consumed and managed by your organisation
- Encourage internal users, such as product managers, to explore the catalogue to identify and find data to reuse in their projects
- Look for opportunities to rationalise datasets if there are multiple, similar entries in the catalogue

5. PROVIDE EFFICIENT WAYS FOR USERS TO ACCESS DATA ON DEMAND

To enable reuse, data needs to be easily accessible, ideally in a 'self-service' environment, allowing users to directly access data or consume it via analytical tools. Approaches might include a mixture of well-documented APIs, a 'data lake'[5], real-time access or bulk downloads. Users may need multiple methods of access to suit different applications.

Recommendation

- Define an informed set of common approaches for sharing data within the organisation that are suitable for use with a variety of tools and needs

[5] A file system containing copies of datasets from across one or more organisations

Why is Thomson Reuters creating a data asset catalogue?

Thomson Reuters is a leading provider of trusted, intelligent information to businesses and professionals in the financial and risk, legal, tax and accounting, intellectual property, science and media domains. The sheer volume and variety of Thomson Reuters content means that it can be difficult to anticipate and judge which content might be useful to a customer. In order to increase the discoverability of Thomson Reuters content, promote self-service and so improve customer access and choice, Thomson Reuters is building a catalogue of financial and risk content assets. Content Kiosk is an app in Thomson Reuters Eikon™ — a platform that combines information, analytics and news from the financial markets. Content Kiosk helps customers, developers and internal users understand not only what data is available, but also discover their meaning, relationships, origin and usage.

Content Kiosk is itself built upon an internal registry of the information assets spanning Thomson Reuters financial and risk offerings called Asset Insight, which is collectively maintained by the groups who own and best understand the assets. Metadata of those assets considered directly reusable by and most valuable to customers are published in the Content Kiosk catalogue; a trade-off between speed of discovery and availability of content. Improvements in catalogue classification and search should remove the need to make an initial selection. The easy-to-navigate catalogue interface presents content from Thomson Reuters and third-party partners. The catalogue is searchable, provides a detailed description of each asset to convey context and includes visualisations and links to the content on the product and the different delivery mechanisms. The use of permanent identifiers allows information to be linked to related material across the Eikon platform, integrating the catalogue with other points of reference and so providing the customer with a more comprehensive understanding of the data, as well as more direct access to the data of highest interest. These features enable Content Kiosk to support a collaborative culture that extends from Thomson Reuters internal product development teams and sales teams to external customers, cultivating a successful data ecosystem.

Effective infrastructures are invisible. They become part of the natural order, only to be remembered when they cease to function as normal.

6. USE COMMON STANDARDS STRATEGICALLY

Effective infrastructures are invisible. They become part of the natural order, only to be remembered when they cease to function as normal. Such infrastructures are governed by a combination of centralised and distributed control and coordination systems, where standards can either travel top-down or bottom-up. Standardisation is the means to unlock the true capability of the network, as different components become interoperable. Standards can apply to the data itself (e.g., formats) and to its context and rights management (e.g., taxonomies, ontologies and levels of access). For example, Thomson Reuters Content Kiosk leverages a content-type and user-type taxonomy to aid in the exploration of its catalogue.

Using common standards allows for modularity, which means that different combinations of known standards can be used to create an effective logical framework customised to facilitate the required operation. The selection of these standards is therefore a key design feature. Those that focus on the ease of use for anyone, and not just those of the local community, should be favoured. The use of standards also allows participants to predict the ability and cost to use data beforehand. Organisations such as the W3C continue to develop standards that address data provenance, rights management and catalogues. Using public resources to describe the context of data, irrespective of the openness of the data, allows for transparency and builds trust.

Recommendations:

- Use common, open standards to help organise how data is managed and published, both internally and with external partners
- Select standards strategically to leverage network effects

Standards that support the creation of shareable data

A number of open standards that support data publishing have been or are being developed. The W3C Open Digital Rights Language (ODRL) standardises the expression of rights information over content such as electronic publications, digital images, software, audio and movies. The Permissions and Obligations Expression Working Group aim to go a step further, to create a flexible and interoperable information model to describe the use of digital content across all sectors and communities. The Data Catalog Vocabulary (DCAT) provides a standard way of defining machine-readable metadata about a dataset, and in doing so increases interoperability between data catalogues. The CSV on the Web Working Group have a similar remit for CSV files on the web; although CSV is a popular format, files differ in the quality and structure of associated metadata making them less portable. PROV is a standard for representing and exchanging provenance information on the entities and activities involved in production, which helps determine data quality. The W3C have also compiled a Data on the Web Best Practices document for supporting a self-sustaining, interactive ecosystem, which can also serve as the basis for enabling robust data publication.

7. DESIGN DATA FOR EVERYONE

When designing data for publication, it is pragmatic to consider it as a contribution to a larger data infrastructure used by many, rather than an isolated item used by one. The need for data that is shareable is driven partly by the unpredictability of future uses and users of that data. The data created must therefore be selected and captured in such a way to achieve a balance between the needs of the application at hand and future usability in order to secure long-term value.

For example, a local requirement may only create the need for train times of departure at the start of the route and arrival at the end of the route as captured by the train driver. It would be more generally useful in a shared data infrastructure if the data published included arrival and departure times at all stops on the route, as captured from the signalling system. Broadening the data usefulness may actually help better satisfy local needs as they change over time and may help reduce future costs.

Recommendations

- Engage with potential reusers of the data to understand their needs, and adapt accordingly
- Consider how the data complements the wider environment and community, rather than concentrate solely on the current use case
- Actively address the compromise between the cost of prior work and the wider benefit of shareable data

8. CULTIVATE AN OPEN, COLLABORATIVE CULTURE

The cultural changes required to encourage the publication and reuse of data are often greater than the technical challenges. A more open internal culture, that embraces a shareable by default ethos, will encourage the creation of data for reuse by others.

Developing successful infrastructure calls for a culture of experimentation and flexibility. However, once established, infrastructures are difficult to change, and initial technical decisions have long-lasting effects. Organisations and communities hence need to achieve a balance that best suits their activities.

Ideally, the costs of creating and maintaining data should be balanced between data providers and consumers. These costs will be offset for both parties by the benefits of increased transparency, visibility and feedback augmenting the value of the data. Specifically for data publishers, and especially for open data publishers, these costs are effectively an investment in the future uses of the data. However, it will be difficult for different groups in your organisation or community to commit to producing shareable data until they can establish the expected cost of doing so. This can be achieved by reaching an agreement on standards and common methods, an action for which some level of coordination and orchestration is required. By analogy, the activities of the Amazon and eBay platforms have drastically accelerated the process for buyers and sellers, by providing standard goods exchange and payment mechanisms which both parties agree on. As the number of participants increase, the network effect makes these standards more attractive. Similar common ground must be cultivated in the selection of standards and methods that make data shareable. Platforms can also facilitate the collaborative managing of data throughout its lifecycle, analogous to how Github facilitates sharing and managing code. This would mean more open and transparent information on data authorities and provenance.

Introducing the technical foundations that enable the sharing and open publishing of data can also lead to, and will support, a culture of collaboration and innovation both internally and externally with selected partners. These frameworks are key for organisations and individuals to realise the true potential value of their data, wherever it lies on the data spectrum.

Recommendations

- Develop and adopt a 'shareable by default' culture that encourages the creation of data for reuse by others and the reuse of data
- Encourage active engagement in and contribution to coordination activities around standards
- Find ways to recognise and reward open innovation within the organisation

Where can I read more?

The following resources provide additional background on some of the ideas explored in this paper:

1. Who owns our data infrastructure? is a short paper from the ODI discussing the growing importance of treating data as an important aspect of society and business infrastructure.
2. Enhancing open data with identifiers describes the importance of identifiers in creating a robust data infrastructure, allowing organisations to link data to other sources to add context, reduce costs and manage complexity.
3. The open data maturity model provides a way for organisations to assess their open data practice. A mature open data organisation will have adopted many of the principles identified in this paper.
4. Read how Thomson Reuters is developing its own open data practice, turning 'open into the new normal'.

AUTHORS:

Tharindi Hapuarachchi (TR), Bob Bailey (TR), Leigh Dodds (ODI),
Andrew Fletcher (TR)

Visit thomsonreuters.com | theodi.org

open data institute

the answer company™
THOMSON REUTERS®