Creating Value with Identifiers in an Open Data World

May 2016

open data institute

the answer company™
THOMSON REUTERS®

Introduction

Thomson Reuters joined the Open Data Institute in March 2014. One of the objectives of that partnership was to work together on collaborative projects that would benefit the wider open data community. This white paper is the first result of our collaboration.

Identifiers are at the heart of how data can be effectively published, retrieved, reused and linked. It's a subject that is fundamentally important to the open data community and to the evolution of the web itself. However, we are at a relatively early stage both in our understanding of the challenges and opportunities that persistent identifier schemes present and also in their adoption for commercial and non-commercial use. Thomson Reuters' experience and insight in this area provides an excellent resource for the community as our understanding and use of identifiers evolve.

This white paper is a joint effort intended to act as a guide to identifier schemes, as well as to start a discussion about how identifiers create value. It can help your organisation understand what you should consider when looking at an identifier scheme and why this is important for data in general and open data in particular. It provides illustrative examples of identifier schemes, many of which are in use by the open data community today. The recommendations of this report should not be taken as a prescriptive set of rules, but rather as a helpful guide that will enable users to unlock the value latent in their data.

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.

Stewart Beaumont
Chief Technology Officer
Thomson Reuters

Sir Nigel Shadbolt
Chairman and Co-Founder
Open Data Institute

Full and summary versions of this white paper can be downloaded from:
thomsonreuters.com/site/data-identifiers/
theodi.org/stories

Management Summary

Identifiers are fundamentally important in being able to form connections between data, which puts them at the heart of how we create value from structured data to make it meaningful. It also turns them into an impediment to creating value when used poorly, and raises a question of how well the identifiers in use today support the goals of the open data movement.

Identifiers are simply labels used to refer to an object being discussed or exchanged, such as products, companies or people. The foundation of the web is formed by connections that hold pieces of information together. Identifiers are the anchors that facilitate those links. The lack of identifiers, or the poor use of them, stifles the power of information gained from linking multiple datasets together. Some of these shortcomings might be overcome using intelligent search and fuzzy matching, but the lower precision of these techniques means that the data never reaches its full potential and there is little incentive to drive improvement of precision over time.

Identifiers are crucial to the process of sharing information, and so fit into many workflows in many different types of workplace. The precision of an identifier fundamentally drives efficiency in a workflow, whether that means referring to a geographic area using a Boundary Line identifier from the Ordnance Survey, or referring to a specific product or resource as part of your supply chain in order to track it without error.

Managing identifiers is easier in a closed system. The web has many advantages, but it presents challenges for identifiers because of its vast scale and their ad hoc usage. Communicating identity — the understanding of what is being described — is essential in conveying the accurate meaning of shared information. This is especially true if the information is shared in machine-readable form, without human intervention. Capturing and representing identity is relatively straightforward in a closed, single-purpose system. However, in an open, multi-purpose environment like the web, which involves many sources of information, it is a more complex process. The scale and ad hoc use of the web means that those who produce and consume identifiers cannot easily coordinate an agreement on the representation and meaning of identity. Much of this coordination relies on the ability and inclination of consumers to look up the definition, usage, validity and equivalence of identifiers. There is no clearly established method for ensuring the communication of identity precisely and at an equitable cost to all.

The complexity and cost of coordinating identifiers raises a particular challenge for open data, whose benefits rest on reuse of information in novel combinations and on low barriers and costs of entry for producers and consumers. As the open data movement is rightly pushing for increasing use of structure and machine-readable data at source, we argue that the challenges of identifiers need to be similarly addressed. Leveraging existing identifiers saves money for each organisation individually by sharing costs, and can be beneficial for big organisations as well as small. For example, by adopting the open music encyclopaedia, MusicBrainz, the BBC saves money overall by redirecting the efforts it would have to take in managing its own identifier scheme towards enhancing an open one.

We can learn from the ways in which identifiers are already being used to unlock the power of open data. This paper draws lessons from illustrative examples and proposes some guiding principles, both for those creating and managing identifier schemes and those who are using them. There are a number of different identifier schemes in use today, using both community-driven and top-down approaches. Through illustrative examples based on the Open Data Institute's experience with data publishing, and perspectives from Thomson Reuters experience in managing its own identifiers, this paper examines why and how the coordination of identity must evolve from being an inherent part of dataset design to being a distinct discipline in its own right.

Recommendations

- **Adopt an open licence.** Administrators or owners of key identifiers in a domain should make those identifiers and any associated descriptive metadata available under an open licence. Using a well-known licence is preferable as it will make the rights and obligations of the consumer easy to understand.
- **Publish useful mappings** data consumers and data publishers have between their own identifiers and external identifier schemes as open data, to simplify data integration for other users.

DATA CONSUMERS:

- **be aware of the design and limitations** of any identifiers they are using, to avoid misinterpreting data
- **avoid misusing and extending identifier schemes that they don't administer**
- **recognise that multiple identifiers exist** for the same entity and either be prepared to manage multiple identities or choose a single authoritative source to align with
- **dereference URLs**, meaning to obtain the latest authoritative metadata associated with an identifier
- **check for any changes** to entities referred to by the identifiers used

DATA PUBLISHERS:

- **ensure that datasets are grounded** with each entity being associated at the right level of granularity with a useful identifier that has associated metadata, e.g., names and labels about their identifiers
- **clearly reference identifier schemes** used in a dataset
- **ensure that any identifiers used in their datasets are compatible with the open licence** applied to the dataset
- **reuse existing identifier schemes** rather than creating new schemes where possible, to encourage convergence within a community

IDENTIFIER PUBLISHERS:

- **provide a reconciliation API** when sharing their own identifiers to allow consumers to match entity names and other characteristics to their identifiers
- **expose documentation** for management of new and existing data frameworks covering the process for assigning identifiers
- **prefer HTTP URLs over other URIs**, ensuring that these resolve to useful metadata about the individual entity
- **ensure that identifiers can reliably be dereferenced** by data consumers and that URL identifiers are c**reated under stable, persistent domain names**
- **provide a stable, highly available means of dereferencing identifiers** that they are committed to providing long term
- **should not delete identifiers once in use** so that objects with only historical existence or objects that have been administratively deprecated can continue to be dereferenceable, returning metadata to indicate their state and, where necessary, linking to any succeeding objects
- **avoid using or creating identifier schemes that allow identifiers to be recycled**
- **provide ways for data consumers to track and synchronise changes** to entities that may affect status or identity, e.g., downloadable daily 'digests' of changes to identifiers and core metadata, http-based dereferenceable identifier URLs or other synchronization options

Open Data & Identifiers

Open data has the potential to help empower citizens, drive transparency in governments, improve supply of public services and drive economic growth by powering new businesses and markets. In large part, delivering on that potential will require the ability to leverage open data from multiple sources and combine it to create unique datasets that can drive applications and support analysis.

Data that is machine readable and published according to open standards is easier to use. Therefore, data publishers must think carefully about how they publish and share data in order to unlock the most value from it. Combining datasets is also easier if the data uses common identifiers, allowing businesses and consumers to identify and link together information about the same objects (locations, products, companies, etc.) published by different sources.

More of this can be achieved by thinking about the identifier scheme to be used from the outset. However, applying an identifier scheme retrospectively can and does offer a business opportunity for others. Adhering to a standard identifier scheme may be the ideal, but it is not always a practical reality. So the opportunity to apply an identifier retrospectively is likely to remain, even once consideration and use of identifiers becomes more common.

Publishing data from a variety of sources under an open licence[1] creates a collection of open data that will support many forms of reuse. Stable, reliable identifiers for companies, documents, locations and products will provide the foundations upon which this collection can grow. Furthermore, consumers will find it easier to combine open data with their own proprietary data, further enhancing the value and use that they can derive from all their data sources.

Building these foundations will require the open data community to work together to overcome a number of challenges that come with creating, maintaining and sharing identifiers. This paper highlights the ways that some of these issues are faced by reusers of open data.

[1] http://theodi.org/guides/publishers-guide-open-data-licensing

Identifiers and Identity

In the small-scale physical world, identity is something that humans intuitively understand because we can easily interact with and distinguish between individual objects. However, in larger-scale environments, particularly where identification becomes a social process, identity can become more nebulous. We quickly find that identifying something — agreeing on its defining characteristics — varies in different domains according to the type of informal (social) or formal (legal and commercial) frameworks that are used.

We can find a simple example of this in geography. Residents in an area might have a colloquial name for a particular landmark or geographic area, e.g., "The Cotswolds."[2] This is a very loose notion of identity as there are no agreed boundaries, but it is sufficient for many ad hoc use cases. Within a legal framework, e.g., to support local government administration, the same geographic area may be more rigorously defined and is perhaps divided into several different regions, each with formal boundaries. Those boundaries may also be drawn differently in order to define electoral districts or to support statistical reporting. In short, the same geographic area ends up with multiple overlapping identities, one for each of those different contexts. All of these different perspectives are valid. The degree of precision with which the identity is defined — and the preferred source of identity — varies based on the specific needs of the individual applications.

The same process applies to more abstract objects. A company may have different identities in different regions based on different legislative and reporting requirements (e.g., company registration numbers). The same novel may be published in different countries under different ISBNs and product codes. People, too, usually have different identities depending on context: Social Security numbers, product logins, employee numbers, and so on.

Identity is always contextual. Different communities will have different rules for assigning identities — different identity models. In today's data-driven world, these differences become important when using data from multiple sources, originating in different contexts. Once an identity has been assigned to something, it needs to be labelled so that it can be referenced within a dataset. The labels, or 'identifiers', are simply tokens used to refer to the object being discussed or exchanged. There are many different types or styles of identifiers. The rules for how an identifier is structured, the means by which it is assigned or validated, and the characteristics of the identity it relates to are often referred to as an identifier scheme. Individual identifier schemes can differ in many ways (see "Anatomy of an identifier scheme").

ANATOMY OF AN IDENTIFIER SCHEME

The following sections highlight areas in which individual identifier schemes may differ. Each of these has practical impacts for data integration and data sharing.

Defining attributes

The selected salient characteristics of an object considered sufficient to establish unique reference between communicating parties.

Syntax

The structure of an identifier, its format and the range of characters used vary from one framework to another. Some identifiers are opaque (e.g., Globally Unique Identifier, GUID) while others are assembled from meaningful components (e.g., Digital Object Identifier, DOI) and in some cases are human readable too (e.g., Reuters Instrument Code, RIC).

Definitions and granularity

The information captured by a framework about the object referred to by its identifiers can differ. Some identifier schemes may be more domain-oriented than others, incorporating defining attributes specific to the intended context. The information considered sufficient to establish an identity and its granularity is also prescribed by the identifier scheme. A higher granularity leads to a higher cost since more information – additional defining attributes – are needed to differentiate the identities.

Scope

Local identifiers may be scoped to a specific database or dataset, whereas a broader identifier might be guaranteed to be unique across many datasets.

Authority

Approaches to assigning identity can be centralised with a given authority assigning its own identifiers (e.g., employee numbers). Others are more decentralised, 'bottom up' approaches that are community-driven (e.g., MusicBrainz).

Discoverability

Some frameworks require all identifiers and their defining characteristics to be deposited in a registry, which allows their pairings to be queried. The publisher or authority for the framework may provide a registry, so consumers would not need to create their own with potentially different results.

Stability

Identifiers are sometimes recycled and reassigned to different objects, inevitably leading to ambiguity and confusion amongst users. C.N used to be the RIC for Chrysler Corp., as listed on the New York Stock Exchange. Following the bankruptcy of Chrysler, C.N was reassigned to Citigroup.

Timeliness and synchronisation

The frequency at which identifier schemes are updated can vary. Some frameworks may push changes to users, while others may require users to work to pull in changes themselves. Synchronisation ensures accuracy and coherence.

Temporality

An identifier can be defined as valid for certain points or periods in time. This is especially useful in managing change of the object being referenced.

Licensing

Identifiers can be subject to different licences, irrespective of the data itself.

[2] An area in the west of England known for its quaint villages and tourist attractions.

Defining your identifier scheme

A key problem for consumers is that different dataset publishers will often use different identifier schemes to identify the same real-world object, for example, the Companies House registration number for a company and its full legal name. Alternatively, different frameworks in different communities may use the same identifier syntax to refer to different objects, for example, HR systems in two different companies using the same name, even though they refer to two different people. Hence, it is necessary to describe the context in which a framework is meant to be used. Frameworks with well-defined syntaxes, such as HTTP Unique Resource Identifiers (URIs) which incorporate the identifier scheme name, allow data reusers to validate data to confirm its context.

Applying your identifiers

Identifiers can be applied in two basic ways. First, data can be published reusing widely accepted pre-existing identifiers to denote the target object. For example, some information published about a musician could be identified with a DBpedia identifier or its corresponding MusicBrainz identifier. Second, if the data uses identifiers created for internal use, or if the data sources being integrated use multiple identifier schemes, these identifiers need to be mapped to others. For example, an internal stock number for paper clips might be linked to an Amazon ASIN for the same item. Similar identifiers can be mapped or linked using various semantic technologies (e.g., sameAs.org). This may be done authoritatively by the group that manages the original identifier or by other users.

Finding their meaning

The meaning of an identifier can be established in a number of ways. A user can look up the defining attributes as provided by the producer of the identifier, or simply adopt its implied meaning by referring to how it is being used by others. In the latter case, meaning can change over time and identifiers can have acquired identities that are different to (or more precise than) that in its original definition. Identifier meanings can also change because the topic or entity it refers to has changed. For example, if company A acquires company B, the identifier for company A now refers to the newly expanded company A, while the identifier for company B becomes obsolete for new data. If the chosen identifier scheme does not incorporate temporal information, users may now choose to replace or equate identifiers in their data for company B as company A, unless they have a particular interest in the two companies pre-acquisition. If there is no temporal information, it can be difficult to ascertain whether an identifier for company A refers to that before or after the acquisition. If the new company is sufficiently different to company A and company B before the acquisition, then neither identifier may make sense for future use. A new identifier may then need to be created.

An added level of complexity is that an identifier might refer to the real-world object (entity) or to a topic signifying description of the real-world object. So, while the identifier for the company B entity may cease to exist after the acquisition, the identifier for the company B topic may still be used in new content, since company B can still be described or discussed retrospectively.

The authoritative identification of an entity and the use of identifiers can have implications because of its contextual nature. For example, in news media, an identifier might refer to a country that has recently been formed, is emerging or even has just been proposed. These entities may not be internationally recognised by one identifier-naming body, but their use may be a necessity within the context of its primary intended use (e.g., a news article). Users must view data through the lens of the identifier scheme they are using, which may or may not align with what is recognised by another identifier scheme.

Organisations and communities define identifier schemes that meet their particular requirements. The design of an identifier scheme always has repercussions for data integration. It is possible that designers may intentionally structure identities in their framework to be imprecise, with the view that users may apply it more broadly to entities in a lexicon. Here, the identifiers may gain an acquired meaning that evolves with use. Other frameworks are targeted at creating a more precise authoritative set of identifiers and may list a larger number of attributes to specify the intended identity, in order to synchronise a clear definition amongst users. The coordination of identity is thus not just an inherent component of dataset design, but should be acknowledged as a distinct discipline in its own right.

The coordination of identity is thus not just an inherent component of dataset design, but should be acknowledged as a distinct discipline in its own right.

Eight Identifier Challenges for Open Data

Open data publishing is still maturing. This means that developers face a number of challenges when attempting to use the data. Many of these relate to the details of the publishing process, e.g., clarity of licensing, choice of publication formats and the regularity of data updates. Clear open licensing, use of structured open data formats and regular updates are the key components of any high-quality dataset. However, there are data-specific issues that can bring fresh challenges, all of which relate to definitions of identity and identifiers.

Each of the following sections presents a specific issue that has been found to re-occur across different open datasets, and is equally valid for closed datasets. The extent to which these issues will be felt will vary depending not only on the individual dataset, but also on the means by or purpose for which the data is consumed. A data consumer using a dataset to perform a one-off analysis or to build a prototype will face different issues to a developer integrating multiple data sources or building a product which requires long-term access to reliable data.

Cataloguing these issues will help data publishers and identifiers' publishers understand and hopefully begin addressing these challenges, resulting in improved data quality for all users. While a publisher needs to focus on ensuring that their own dataset is published in a sustainable way, ensuring successful and widespread use of that dataset by consumers means it is essential to consider the context(s) in which that data will be reused. This often means considering how the data could or should link to other sources. Similarly, if reusers begin sharing solutions to the challenges they face, the community can develop tools and techniques to help simplify open data integration.

... data may be ungrounded simply because the publisher has overlooked the need to publish and share an identifier that almost certainly exists in the database from which the data was originally sourced.

1. DATA IS UNGROUNDED

Without identifiers, data is ambiguous and places a burden on data consumers.

A 'grounded' dataset is one in which every entity referenced in the data is associated with a suitable identifier. Access to an identifier that is explicitly managed allows reusers of the data to locate data about those entities, clarifying their meaning and role in the context of that dataset. Identifiers also make it easier to discover more data published about the same entity. Unfortunately, many datasets are either completely or partially 'ungrounded'. Instead of including a useful identifier, it may only include a name or unexplained label for an entity.

For example, election results may include only the names of electoral districts and political parties, but not an official identifier. Given that political parties often use similar names, the lack of an identifier can make it hard to determine whether a variation in a name is a mistake or a reference to a completely different party

Ungrounded data is particularly common in datasets that are published from spreadsheets that have been designed to be human- rather than machine-readable. The data may be tidily and clearly formatted but lack connections to useful identifier schemes. In this scenario data may be ungrounded simply because the publisher has overlooked the need to publish and share an identifier that almost certainly exists in the database from which the data was originally sourced.

However, in some cases it may be that the data publisher does not have an identifier for all entities. A spreadsheet of financial transactions, for example, might have an identifier for each transaction, but the publisher may not have collected or assigned a unique identifier for all of the companies involved in those transactions.

Ungrounded data places a significant burden on its consumers to identify the entities to which the data actually refers. It also makes it very difficult to accurately match entities across datasets, which is essential when connecting or enriching data using multiple sources. This lack of clarity and the need to 'fuzzy match' data post-publication will introduce mistakes which impact on correct aggregation and analysis.

Recommendations:

- **Data publishers should ensure datasets are grounded,** with each entity being associated at the right level of granularity with a useful identifier that has associated metadata, e.g., names and labels about their identifiers

Illustrative Example 1:

OPENFDA

The OpenFDA APIs provide details of adverse drug events and recall notices. The data exposed by the API is grounded in a number of identifier schemes, including standard drug codes and ingredient identifiers.

Domain: **Health**	Type: **Data Infrastructure**
Key Challenges Addressed • Grounding of data • Rationalising multiple identifiers	

The Food and Drug Administration (FDA) in the US launched openFDA[3] in 2014, an API enabling easy access to their publicly available public health datasets. The platform is focused on adverse events and product recalls datasets for FDA-regulated drugs, foods and devices, and structured product labelling data for FDA-regulated human drug products. Although this data has been publically available, it has been difficult to use. For example, the platform brings together information scattered in over 3 million adverse-events reports, making them more accessible and queryable.

OpenFDA is geared towards facilitating application, mobile, and web developers and researchers to use FDA data in their work. The platform has great potential, from applications for professionals in the legal domain, looking for evidence of a drug's adverse effects, to mobile apps that allow consumers to check for these and product recalls at the point of sale.

All data available through the openFDA platform is released with a Creative Commons CC0 Public Domain Dedication licence. This means that users can distribute, modify and work on the data even for commercial purposes without seeking permission.

An important part of the initiative has been the harmonisation of drug identifiers across their different datasets, making it easier to search for and understand products in context of the datasets presented. These integrations require an exact match, however, and so it is not a straightforward process. For example, the platform extracts adverse-events reports submitted by medical professionals, which sometimes includes misspelled drug names resulting in mismatches.

[3] https://open.fda.gov/

2. LACK OF RECONCILIATION OPTIONS

Inability to resolve a name or code into an identifier introduces additional overheads.

Ungrounded data means that developers often need to resolve a name or label into a more stable identifier. Names of objects may change, so using a stable identifier is a better foundation for a useful dataset. This is often the first step in trying to link together datasets that have been published without common identifiers. For example, a data consumer may need to turn the name of an electoral district into the identifier for that specific administrative region. Similarly a company name may need to be resolved into a stable identifier. This process is commonly known as 'reconciliation'.

Reconciliation either requires access to bulk downloads of identifiers and their related metadata, or the availability of a query API that will allow developers to look up the identifier for a specific name or set of metadata elements. Without access to these services, data reusers must take on the added burden of trying to manually build lists of correspondences between names and identifiers and maintaining them across time, which is additional work and leads to differences of understanding between reusers.

Recommendations:

- Identifier publishers should provide a reconciliation API to allow consumers to match entity names and other characteristics to their identifiers

Illustrative Example 2a:

OPENCORPORATES AND OPENLEIS

OpenCorporates provide a reconciliation API to support data consumers in matching company to legal company entities

Domain: **Business/Legal**	Type: **Dataset & Services**
Key Challenges Addressed	
• Open Identifier Scheme • Reconciliation to identifiers • Resolvable identifiers • Rationalising multiple identifiers	

The largest openly licensed database of companies across the globe, OpenCorporates[4] aims to collect unique identifiers for all corporate entities. It was set up in 2010 by Chris Taggart and Rob McKinnon, who were driven by a lack of clarity in global government data relating to companies. They found that the data was often inaccurate, incomplete, out of date and sometimes duplicated and unlinked across government registers. The project collects basic, essential information on companies and government data that relates to them in an effort to increase the understanding and transparency of company data.

The team uses a number of data sources in addition to company registers. These include a wide array of national and global datasets such as the latest world trademark register, the latest data from the US's Central Contracting Registration system and the daily London Gazette, where official insolvency notices are published. At the time of writing, OpenCorporates had amassed information on over 78 million companies in almost 90 jurisdictions. Users can access this information via the OpenCorporates API or Google Refine reconciliation services.

OpenCorporates uses a 'share-alike' licence, where the data is free for anyone to use as long as any product of that data is also kept open. Users who don't want to be restricted by the share-alike licence pay OpenCorporates a fee for the privilege.

The release of government data has been one of their main issues, as governments look to gain an income from company registers. However, there are signs that these attitudes may be changing. In the UK, Companies House recently announced that it will be offering all its digital data free of charge, becoming the first country to establish an open register of business information.

In 2013, OpenCorporates launched a sister website, OpenLEIs[5], which is an interface on the Global Legal Entity Identifier System (GLEIS). Also known as the LEI system, GLEIS provides an open, persistent identifier for each corporate entity. Existing corporate and legal identifiers are subject to licensing restrictions which make publishing and using data with these identifiers a difficult process. The system is an initiative by the G20 and the Financial Stability Board for use in the financial markets, where issues using company identifiers and the lack of information on links between corporate entities have become more significant in the wake of the financial crisis. It is thought that entities will be issued by local operating units, who will be coordinated by a central operating unit at the Global Legal Entity Identifier Foundation. The foundation is currently being set up by the Financial Stability Board. The Regulatory Oversight Committee will in turn oversee the LEI system, to ensure that "reliable, while flexible, operational principles and standards [are] applied to the origination and maintenance of an LEI."[6]

The user-friendly OpenLEIs interface allows you to search entities by name, address, legal form, registering body or a combination of these queries. Every LEI has a permanent URL and links to OpenCorporates where applicable, providing more detailed information, all of which are available as open data.

It remains unclear if the identities defined by OpenCorporates and OpenLEIs identifiers are truly equivalent to those defined by existing identifiers that are widely used in industry today. Unless OpenLEIs are used globally, and within organisations, internal or licensed identifiers will still need to be mapped to open ones.

[4] https://opencorporates.com/
[5] https://openleis.com/
[6] http://www.leiroc.org

Illustrative Example 2b:

OPENCALAIS

Launched by Thomson Reuters in 2008, OpenCalais enables users to quickly and automatically tag people, places, companies, facts and events in any given textual content, improving its interoperability on the web.

Domain: **All**	Type: **Infrastructure**
Key Challenges Addressed • Reconciliation to identifiers • Rationalising multiple identifiers	

The service enables users to quickly and automatically tag people, places, companies, facts and events in any given content, improving its interoperability on the web. Launched in 2008 by Thomson Reuters, OpenCalais[7] enriches user content with rich semantic metadata, expanding the semantic web.

OpenCalais employs Natural Language Processing (NLP), machine learning and other techniques to extract named entities, facts and events from user-submitted content and tags them with metadata.

This allows users to build graphs or networks linking content to real-world objects and topics, improving site navigation, and providing contextual syndication and the ability to organise and analyse content. When an entity is extracted, Calais returns an identifier (a HTTP URI) that can be dereferenced to provide useful information and links to other relevant data and web assets. Calais links to a variety of assets including DBpedia, Wikipedia, Freebase, Reuters.com, GeoNames, Shopping.com, IMDB and LinkedMDB.

The OpenCalais API is accessible as a web service and is free for both commercial and non-commercial use. The initiative fosters a community of developers to help build applications and tools for everyday users. Calais Marmoset makes websites ready for intelligent search. For example, it can determine whether the word "Washington" on a website refers to the city, the state or the person. Search engine crawlers that use this metadata are then better able to provide users with targeted results. Calais Tagaroo is a WordPress plug-in for bloggers which suggests tags and Flickr images to embellish a post. Calais also offers a submission tool where users can submit files for semantic analysis.

3. LACK OF IDENTIFIER SCHEME DOCUMENTATION

Lack of clarity of how identity and identifier models are created, maintained and intended to be used hinders successful reuse.

Datasets often lack useful documentation for developers. Documentation is essential to ensure that data can be correctly processed and analysed. A good set of documentation should answer:

- What identifier scheme(s) are used in the data?
- If the dataset includes new custom identifier schemes, how is that identifier scheme defined, e.g., what is the syntax for the identifiers? (See 'Anatomy of an identifier scheme'.)
- How are identifiers assigned? E.g., what is the 'identity model' used to assign identities to things in this dataset? Which types of things are relevant and included and which are considered irrelevant and not included? How granular is the model?
- When might a new identifier be assigned to an existing resource? What are the change control processes?
- What is the life cycle of the identity and how is change communicated and synchronised between those using that identity?
- Are identifiers guaranteed to be stable, or might they change, be reused or be re-assigned?
- Are there relationships between identifiers, e.g., are they organised into a hierarchy? Are there explicit relationships with other types of entity: is this part of a broader information model? (For example, financial securities are issued by organisations: an identifier scheme for securities might be explicitly connected to an identifier scheme for organisations.)

This type of information is essential for data integration as it helps to:

- support data validation: e.g., to validate identifiers against their expected syntax
- communicate semantics: e.g., clarify what types of entity are being described or referenced
- avoid misinterpretation, e.g., treating a key as unique and stable when in fact it isn't;
- bring clarity to what is being described, e.g., define what type of administrative districts are referenced

Some identifier schemes are intended to be used as a standard within a community. This requires that the framework is properly documented so that the identifiers can be reused effectively, e.g., to publish new data as 'annotations' against existing identifiers.

Thomson Reuters Perspective

When creating identifiers, it is advisable to have a single method for producing identifiers regardless of the type of entity they refer to. Having different methods for generating identifiers for people, companies and places increases the amount of work that needs to be done both in producing and understanding them.

Data identifiers should be opaque (not human-readable) in order that they cannot carry information that might be interpreted separately from their actual corresponding identity. Human-readable symbolic names should be separately registered and mapped to the data identifiers (thereby being convertible and at the same time creating a thesaurus of equivalent symbolic names).

Recommendations:

- Data consumers should ensure that they are aware of the design and limitations of any identifiers they are reusing, to avoid misinterpreting data
- Data publishers should clearly reference identifier schemes used in a dataset
- Identifier publishers should expose documentation of new and existing data frameworks covering the process for assigning identifiers

[7] http://opencalais.com/

REUTERS INSTRUMENT CODE (RIC)

The Reuters Instrument Code (RIC) is a well-known and documented human-readable identifier for financial instruments and indices. It has been doing 'double duty' as a data identifier too — which over time has put strain on its ability to satisfy the combined needs of people and machines.

Domain: **Business**	Type: **Dataset & Services**
Key Challenges Addressed • Documented identifier scheme • Resolvable Identifiers • Reconciliation to identifiers	

The Reuters Instrument Code (RIC) stands apart from the other examples in this white paper since it is not under an open licence and consequently is not in use by the open data community. However, as a popular Thomson Reuters identifier, it still provides useful insights for the challenges identified in this paper. Furthermore, having being designed in the 1980s, long before the recent awareness and release of open data, it provides an interesting counterpoint to more recent identifier schemes.

The RIC was created to be a logical and intuitive symbolic name for financial instruments and indices, designed to be used by clients' users for intuitive querying and navigation of Thomson Reuters sourced data.

The RIC is made up of the security's ticker symbol, followed by a period and an exchange code based on the name of the stock exchange that uses that ticker. For example, IBM.N refers to IBM traded on the New York Stock Exchange, and IBM.L refers to the same stock traded on the London Stock Exchange. The same ticker symbol can refer to different securities within an exchange as ticker symbols are often reused. The RIC was originally designed to include a component unique to each company, and intended to be freely available. These ideas were however vetoed at the time.

The human-readable format of the identifier has meant that users often simply adopt a RIC without dereferencing it, on the assumption that it refers to a definition that they interpret or remember from its syntax. These practices have inevitably led to confusion amongst users. For example, the RIC 'C.N' at one time referred to Chrysler Corp., but when that company was bought by Daimler, 'C.N' was reused to refer to Citibank Corp, simply because that company adopted the New York Stock Exchange ticker 'C'. Many users continued to assume the Chrysler Corp. meaning when using the RIC for communication and data lookup, creating ambiguity at the time.

In 2009, the European Commission opened an antitrust investigation on concerns that Thomson Reuters may have abused its market position in the real-time consolidated datafeed space by preventing customers from using RICs to source data from competitors. According to the Commission, RICs had become embedded into customers' applications; however, when customers chose to switch to a competing real-time consolidated datafeed product, they had to remove the RIC from those applications and could not use them to source the third-party data. As a solution, Thomson Reuters offered to create a new licence that would allow previous customers of its real-time consolidated service to continue using the RIC in their applications post-switch, and the investigation was dropped without any adverse finding against Thomson Reuters. The Commission Decision is being challenged by another data vendor, Morningstar.

SHOULD IDENTIFIERS BE HUMAN-READABLE?

Web identifiers — URIs — are represented and communicated as a string with a specific syntax, such as the HTTP URL **http://thomsonreuters.com/financial/equities-and-derivatives/** for a web page or **https://opencorporates.com/companies/ca_pe/144369** as an identifier representing a real-world object.

A common design choice is whether it is desirable that the URI string should be able to be interpreted for its meaning by anyone reading it — should the identifier be 'human-readable'? Or is 'human readability' something that may work against the intended purpose of the identifier — so should it be opaque? This decision hinges on:

1. whether a person will ever actually see the URI (or need to see the URI)
2. whether any interpretation by a person is likely to be considered useful in respect to the thing being identified: does it tell the reader anything reliably useful
3. whether interpretation by a person is likely to agree with or differ from the actual meaning of the thing being identified as understood by machines using it

If the identifier is not intended to be precise as to the meaning of the thing identified — i.e., different users of the identifier may draw different conclusions as to the meaning of the thing identified — then human readability may be beneficial in conveying something about the owners' intended use of it.

If the identifier is intended to be a proxy for a very specific meaning, human readability of the identifier raises the probability that a person or programme interpreting it will reach a different conclusion of its meaning than the machines using it, especially if the specific meaning might change over time (most real-world things change to some degree across time, such as geopolitical boundaries, a company's legal status or regulatory rules). Adjusting the identified data to represent changed meaning is comparatively easy; adjusting the identifier to match is hard, since it requires communication and resynchronisation of identifier use, and risks broken references.

For example:

- **http://thomsonreuters.com/financial/equities-and-derivatives/** identifies a page of description whose 'meaning' is not singular and must be interpreted by reading or extracted by analytics. The string gives some idea of the owners' intended use but nothing else.
- **https://opencorporates.com/companies/ca_pe/144369** identifies a very specific real-world object whose meaning can only be fully communicated with a range of data points:
  - 'THOMSON REUTERS (FINANCIAL & RISK) CANADA'
  - Company Number 144369
  - Incorporation Date 10 July 2012
  - Jurisdiction: Prince Edward Island (Canada)
  - Status: active

To be correspondingly informative about meaning, the identifier would have to have contained all that information. That would then lead to difficulties should any of that data conveying meaning change or be corrected, undermining the identifier's stability.

- **http://dbpedia.org/page/Thomson_Reuters** identifies a fairly broadly named concept: Its meaning is not intended to be singular. It acts as a point of aggregation for facts relevant to the broad concept, which its identifier usefully communicates to people. However, it may not be useful to machines testing for exact matches of meaning.

4. PROPRIETARY IDENTIFIER SCHEMES

Closed identifiers hinder the growth of the open data commons.

Some communities use common identifiers to help simplify data exchange. These identifiers may be from a formally standardised framework or, more commonly, may have just become the de facto standard because of their ubiquity. In many cases the identifiers and any related metadata captured by the scheme are the intellectual property of a single organisation. If this is the case, the user needs to consider the licence the data is under, in order to distinguish between, for example, an independent industry body like ORCID (Open Researcher and Contributor ID) and an organisation managing identifiers as a closed, proprietary asset.

While this doesn't always impact data integration of closed data, it does have a severe impact on the open data commons: organisations cannot publish any open data that uses proprietary identifiers. Inclusion of such identifiers in an open dataset "poisons the well", making it impossible to apply an open licence to what would be considered a derived dataset. Open data and closed identifiers don't mix.

An example is the Unique Property Reference Number (UPRN), a 12-digit identifier assigned to every building and plot of land in the UK. This identifier is therefore an essential part of the UK digital infrastructure and is used in many databases across a variety of organisations. However, the licensing terms for the UPRN prohibit its use without the purchase of the commercial product from which it ultimately derives. This means that no open dataset can include a UPRN.

The UPRN example illustrates a failure to distinguish between the utility of using an identifier, e.g., to reference it in a derived dataset, and the publishing of all the valuable primary data associated with that identifier. The more an identifier is used in other datasets, the more valuable the primary data becomes. The source of value is from the network effect of using an identifier, and the ability to exchange it for useful data: It is not the identifier itself which has value.

Closed identifiers, therefore, seriously impede the growth of the open data commons, as both data publication and sharing is severely restricted. Organisations either have to publish 'ungrounded' data, which lack the key identifiers required to join datasets, or must create and maintain new identifier schemes, which may take more time to be widely adopted.

Data publishers should consider the licensing terms associated with the identifiers referenced in their datasets. Because of the network effects mentioned above, rather than being a burden, it may well be in their own interest to choose an open licence for the identifier. Where they are the authority, then the identifier scheme should be not only well documented but also openly licensed. If a data publisher is not the authoritative source for an identifier, then care must be taken to ensure that the identifiers are open data compatible, or that a substitute identifier scheme is used. Otherwise, datasets that use those identifiers cannot be classed as open data.

Recommendations:

- Adopt an open licence. Administrators or owners of key identifiers in a domain should make those identifiers and any associated descriptive metadata available under an open licence. Using a well-known licence is preferable, as it would make the rights and obligations of the consumer easy to understand
- Data publishers should ensure that any identifiers used in their datasets are compatible with the open licence applied to the dataset

Open data and closed identifiers don't mix.

Illustrative Example 4a:

NHS ORGANISATION DATA SERVICE

The NHS Organisation Data Service publishes standard codes that identify both institutions and individuals who provide health and social care services, or interact with the NHS in some form. The data is openly licensed, supporting use of the identifiers within the NHS and also by third parties.

Domain: **Health**	Type: **Dataset & Services**
Key Challenges Addressed • Grounding of data • Documented identifier scheme	

The NHS Organisation Data Service[8] (ODS) operates as part of the Health and Social Care Information Centre (HSCIC). Its role is to publish standard codes that identify both institutions and individuals who provide health and social care services, or interact with the NHS in some form.

The codes published by the ODS form part of the NHS data standards and are used through the organisation. The codes are associated with metadata that describe the relationships between different care providers, their geographical location, etc. As well as administering the identifier scheme, the ODS are responsible for publishing this data to all interested parties both within and outside the NHS.

The ODS codes, along with many other datasets from the NHS, including key performance statistics, are published as open data under the UK Open Government Licence. The codes act as a common reference point that enables the integration of a number of datasets from across the NHS. The identifiers, being open and clearly maintained, can be reused by third parties wishing to contribute additional metadata about NHS organisations.

[8] http://systems.hscic.gov.uk/data/ods

Illustrative Example 4b:

ORCID

ORCID provides an open identifier scheme for use within the publishing industry. It helps solve the challenges faced with data integration, which historically was hampered due to either the lack of identifiers or the use of proprietary frameworks.

Domain: **All**	Type: **Data Infrastructure**
Key Challenges Addressed • Open identifier scheme • Reconciliation of identifiers	

The Open Researcher and Contributor ID (ORCID)[9] is a community-driven initiative to create a registry of unique stable identifiers for researchers, which facilitates a transparent method of linking research activities and outputs across disciplines, institutions and geographic boundaries. Thomson Reuters is a founding member and has worked with the organisation since its start in 2009.

ORCID provides an API to connect external applications to their registry. Several organisations are using, or have started to integrate, ORCID identifiers in their work, including publishers and research funding bodies allowing them to track the outputs and outcomes of research that they have funded. It also links to other popular frameworks such as Scopus, ResearcherID (powered by Thomson Reuters) and Google Scholar.

A number of features are available for free. Individuals can register, maintain and share their ORCID identifier and record data, and can search ORCID records and view public data. ORCID also releases a file of public information annually, under a Creative Commons CC0 1.0 Universal Public Domain Dedication licence. Members and subscribers benefit from Member APIs that provide access to both read and write data in existing ORCID records and to generate new records for individuals at member organisations.

ORCID does not prescribe to a strong notion of identity, however — researchers can have more than one ORCID. Often the system pulls in duplicate entries of an author's work, as differences in any component of the citation (such as external link or page number) are marked as distinct works.

The more an identifier is used in other datasets, the more valuable the primary data becomes

[9] http://orcid.org/

5. RATIONALISING MULTIPLE IDENTIFIERS

Even when communities converge on standard identifiers, entities will often have multiple identifiers from different sources.

While some organisations exchange data using shared identifiers, it is much more common to find that different organisations have assigned their own identifiers to the same entity. This is also a problem within individual organisations where there is little or no coordination in data management practices between departments.

The first step in combining data from multiple sources is often the creation of lookup tables that list equivalent identifiers from each of the different sources. These tables are often compiled over time and require continual maintenance to ensure that they stay current. They are also a frequent source of error: It may not always be clear when two different identifiers actually do refer to the same entity or whether, because of the use of different identity models, they refer to separate resources.

Maintaining and defining equivalence is often best done 'at source', e.g., each organisation clarifies which third-party identifiers are equivalent with its own. But this is costly and often difficult to get right, even for small number of mappings. It also doesn't scale when there are very large numbers of organisations exchanging data, as is the case with open data.

These costs encourage convergence on the use of standard identifiers. Top-down approaches, e.g., defining new standard identifier schemes within a community take considerable time and effort to achieve agreement and then widespread adoption. Top-down agreement may also be difficult to achieve across communities. In contrast, bottom-up approaches happen when organisations choose to reuse existing identifier schemes, rather than creating new identifiers. This approach requires much less co-ordination, allowing authoritative sources of identity to appear over time. However, this still requires users to adopt the same identity model as the owner of the identifier, albeit in a more distributed way, rather than picking something that is 'close enough'. As a benefit, investing time up front in carefully choosing an existing identifier reduces the burden on reusers and consumers.

Regardless of the above, we expect that dealing with multiple identifiers will remain a challenge in the long term. Owners of identifier schemes might usefully lower the burden on consumers by providing extensive cross-mapping to equivalent frameworks. Consumers may need to adopt more fluid ways to layer together datasets that use different approaches for modelling the world. The circumstances in which it is important to define strong equivalences, and those in which looser notions of equivalence are important, will depend on the individual application.

Recommendations:

- Both data consumers and data publishers should publish any useful mappings they have between their own identifiers and external identifier schemes as open data, to simplify data integration for other users
- Data consumers should avoid misusing and extending identifier schemes that they don't administer
- Data consumers should recognise that multiple identifiers exist for the same entity and either be prepared to manage multiple identities or choose a single authoritative source to align with
- Data publishers should leverage existing identifier schemes where possible to encourage convergence within a community

Thomson Reuters Perspective

Cross-mapping separately organised identifier schemes results in lower overall information coherence than shared working to authoritative identity schemes. Whether this is good or bad depends on the intended use of the resulting information: Audience preference for authoritative information will prefer greater coherence, while audience preference for information that evolves naturally with actual use will prefer cross-mapping frameworks that have 'acquired meaning'.

Illustrative Example 5:

SAMEAS.ORG

SameAs.org provides APIs that allow developers to discover equivalent identifiers published by many different sources.

Domain: **All**	Type: **Data Infrastructure**
Key Challenges Addressed • Rationalising multiple identifiers	

SameAs.org[10] is a service whose aim is to address some other aspects of the challenges caused by having multiple identifiers for the same object. The service collects together equivalence relationships defined between many different linked open data datasets, wrapping that information in APIs that allow those equivalences to be discovered.

For example, SameAs.org can identify which other identifiers are equivalent to the DBpedia identifier for Edinburgh. Having discovered those identifiers, developers can use those links to find additional metadata about Edinburgh from a richer variety of sources. Data that uses these equivalent identifiers can be easily merged into existing sources.

However, SameAs.org recognises that because defining equivalences is not an exact science, it allows developers to query for only those mappings that come from specific sources. For example, the identifiers that the British Library considers to be equivalent with its own may be a higher-quality set of mappings than those defined by third parties. This allows different perspectives on equivalence to be layered on top of source data, providing a more flexible way to integrate data.

[10] http://sameas.org/

6. INABILITY TO RESOLVE IDENTIFIERS

Identifiers which are not integrated with the web require extra infrastructure to support data discovery and determine authority.

Identifier schemes often associate identifiers with at least some minimal metadata, e.g., one or more labels that provide a name or title for the identified object and some administrative metadata indicating when the identifier was created and other life-cycle metadata. In the case of geographic entities, the identifier authority may also define a boundary, for example, as a connected set of latitude and longitude points that enclose the area.

It is often useful to be able to look up (or resolve, or 'dereference' an identifier to obtain the core metadata with which it is associated. This functionality should be through an appropriate API but in many cases will also benefit from having a more human-readable user interface, such as through a web-based lookup. This supports a number of use cases, including:

- checking if the identifier is valid and still in use
- finding a display name or description to help build a user interface around a dataset
- discovering pointers to additional useful datasets
- establishing equivalence with other identities (i.e., other identifiers that genuinely refer to the same real-world object)

Identifier schemes that are based on simple coding framework, e.g., simple alphanumeric sequences like the ISSNs, ISBNs and GUIDs, don't provide a means to resolve those identifiers into useful data. Without additional documentation defining the source of the identifier, it can be difficult to determine the authoritative source. Developers must therefore discover this additional context for themselves, which may be difficult if documentation is lacking. This contributes additional upfront overhead when attempting to aggregate data from multiple sources.

Contrast this with frameworks based on URLs that include a built-in way to retrieve metadata about the identified resource: by making a simple HTTP request to the individual URL.

This is the foundation for the concept of linked data: the ability to easily exchange any identifier for additional, trusted, contextual data published by the identifier authority. This metadata can be obtained on demand as new identifiers are discovered. Datasets that use URLs are integrated with the web; following links in the dataset can greatly simplify the process of discovery of relevant extra data.

URL-based identifiers also have authority built in: The URLs are based on an internet domain that will be registered with an easily identifiable company or organisation.

Recommendations:

- Identifier publishers should prefer HTTP URLs, over other URIs, ensuring that these resolve to useful metadata about the individual entity
- Data consumers and reusers should ensure that they dereference URLs to obtain the latest authoritative metadata associated with an identifier

Datasets that use URLs are integrated with the web; following links in the dataset can greatly simplify the process of discovery of relevant extra data.

Illustrative Example 6 and 7:

DBPEDIA

DBpedia provides a linked data interface that exposes identifiers and metadata for all of the entities described in Wikipedia.

Domain: **All**	Type: **Dataset & Services**
Key Challenges Addressed • Reconciliation to identifiers • Resolvable identifiers • Rationalising multiple identifiers	

DBpedia[11] is a community-driven project that aims to provide a machine-readable view of the metadata and relationships captured in Wikipedia. It uses a crowd-sourced set of mappings to convert structured text from Wikipedia into machine-readable data.

The breadth of Wikipedia means that it has pages for many different objects. These pages are used to create a unique identifier for that object in DBpedia. Through this process, the DBpedia dataset has quickly become a 'Rosetta Stone' that connects together many datasets in the Linked Data Cloud.

Individual data publishers have defined equivalences between their URIs and those defined by DBpedia. By using this common reference point, data publishers can concentrate on managing one set of identifier equivalences: from their data to DBpedia, rather than to many different sources. Yet, in combination, these 1:1 mappings combine to provide links that integrate many different datasets. This helps to scale the creation and management of definitions equivalence across the user community. This in effect confers 'acquired meaning' on the DBpedia identity since while publisher A might believe their ID is the 'same as' a DBpedia equivalent, their interpretation may differ from another, B, who also believes there is equivalence between their ID and DBpedia. This loose equivalence is good enough for navigation by humans, facilitating easy recall of information, but can cause problems for machines that require higher levels of precision.

There are other challenges to be faced. For example, as Wikipedia is edited by its community, it is possible that the meaning and metadata associated with a DBpedia resource will change over time. Wikipedia pages may also be deleted at any time. This raises questions about the stability of DBpedia URIs over the long term.

[11] http://dbpedia.org/

7. FRAGILE IDENTIFIERS

Creating reliable identifiers strengthens the data commons, but comes at a cost.

The use of HTTP URIs as identifiers does have its own downsides, however: If the expectation is that a URL should be resolvable into open data about the identified resource, then the authority must deliver that service and ensure that those identifiers will remain accessible over a long period, ideally permanently.

While an unresolvable URL may still be perfectly useful as a simple identifier, the benefits of being able to easily access metadata are lost. This can be mitigated by publishers providing bulk downloads of data. As well as addressing resolution issues, bulk downloads can also support archiving and local processing of data.

This places additional requirements on data publishers to ensure that they are delivering these identifiers as part of a stable, reliable service. Identifiers that are intermittently available may not be reused by the wider community if there are concerns about their reliability. The costs of maintaining URLs are also higher, especially for highly used identifiers, which may result in high volumes of usage to a data publisher's web servers.

Recommendations:

- Identifier publishers should ensure that identifiers can reliably be dereferenced by data consumers and that URL identifiers are created under stable, persistent domain names. Identifier publishers should be prepared to provide a stable, highly available means of dereferencing identifiers that they are committed to providing long term
- Identifier publishers should not delete identifiers once in use. Any objects that have only historical existence or objects that have been administratively deprecated should continue to be dereferenceable, returning metadata to indicate their state and, where necessary, linking to any succeeding objects

8. IDENTIFIER RECYCLING AND EVOLUTION

Unstable identifiers and changing notions of identity create challenges for data integration in the long term.

Another form of identifier instability comes from using the same identifier to mean different things at different points in time. Ideally, a stable identifier would always uniquely refer to the same entity in a stable state. However, there are several circumstances when existing identifiers may end up referring to different entities, e.g.:

1. when the identifier is reused to identify a completely different entity. For example, while ISBNs are not meant to be reused, there are circumstances where two different books have been given the same ISBN.
2. the identifier refers to an entity whose identity has changed over time. For example, a UK postcode might have addresses added or removed resulting in changes to its boundary.

While the first example may be easy to spot, the second is more subtle. Depending on how the data is being used, the changes may not be significant. But without a clear description of the entity and how its definition has been changed, it may be difficult to tell.

Recommendations:

- Identifier publishers should avoid using or creating identifier schemes that allow identifiers to be recycled
- Identifier publishers should provide ways for data consumers to track and synchronise with changes to entities that may affect status or identity, e.g., downloadable daily 'digests' of changes to identifiers and core metadata, HTTP-based dereferenceable identifier URLs or other synchronization options

Illustrative Example 8:

MUSICBRAINZ (ODI)

MusicBrainz is an example of how a community can work together to overcome the challenges of fragile identifiers. By acting as a clearinghouse for users who curate the data and working with large partners such as the BBC, while including examples of duplication, the quality of the identifiers and associated metadata can consistently improve over time.

Domain: **Media**	Type: **Dataset & Services**
Key Challenges Addressed • Reconciliation to identifiers • Resolvable identifiers • Rationalising multiple identifiers • Open Identifier Scheme	

MusicBrainz[12] is a community-maintained database of music metadata. It catalogues metadata about artists, albums, tracks, musical works, releases and places — all of which are associated with a unique MusicBrainz identifier which is exposed as a URI that can be resolved to retrieve the metadata about the entity.

The service provides a rich set of APIs to help reconcile, e.g., artist names, to MusicBrainz identifiers. It also collects a rich set of equivalences between its identifiers and those generated by other systems, including Wikipedia.

All of the core metadata has been published under an open data licence for many years, resulting in its incorporation into many products. The BBC Music website uses MusicBrainz as its primary source of both data and identifiers. Rather than create its own database and identifier scheme, the BBC editorial staff contribute to MusicBrainz as a shared resource. Internal data can be easily linked to the open data through the use of the common MusicBrainz ID. Additional metadata, e.g., artist biographies from Wikipedia, is sourced through the equivalent identifiers contained in the database.

MusicBrainz therefore acts as a clearinghouse in which both individuals and organisations can collaborate on curating a high-quality dataset and set of identifiers for the music domain.

[12] http://musicbrainz.org/

Summary

The design and use of successful identifier schemes involves a mix of social, data and technical engineering. Ideally, identifiers should be stable, discoverable and clearly defined in terms of their scope and potential for change over time. Identifiers and their core metadata should be open and free to use by any third party.

Successful frameworks will also balance the needs of the data publisher and the data consumer. Lowering the degree of coordination that publishers and consumers are required to maintain will improve the ease with which data can be shared and reused. The specific challenges and illustrative examples explored in this white paper have addressed some of these issues. Whilst equally valid for closed datasets, most of the examples have been chosen with a focus on the recommendations and how their impact will relate to the growth of the open data commons.

Successful frameworks will also balance the needs of the data publisher and the data consumer. Lowering the degree of coordination that publishers and consumers are required to maintain will improve the ease with which data can be shared and reused.

AUTHORS:

Open Data Institute — Leigh Dodds, Georgia Phillips

Thomson Reuters — Tharindi Hapuarachchi, Bob Bailey, Andrew Fletcher

Visit thomsonreuters.com | theodi.org

open data institute

the answer company™
THOMSON REUTERS®