Creating Value with Identifiers in an Open Data World

Management Summary

October 2014

open data institute

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.

James Powell
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

AUTHORS:

Open Data Institute — Leigh Dodds, Georgia Phillips

Thomson Reuters — Tharindi Hapuarachchi, Bob Bailey, Andrew Fletcher

Visit thomsonreuters.com | theodi.org

open data institute

THOMSON REUTERS™