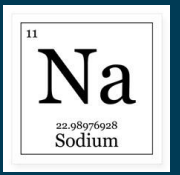
KC2 webinar: DataCite DOIs

Martin Fenner, Team Sodium



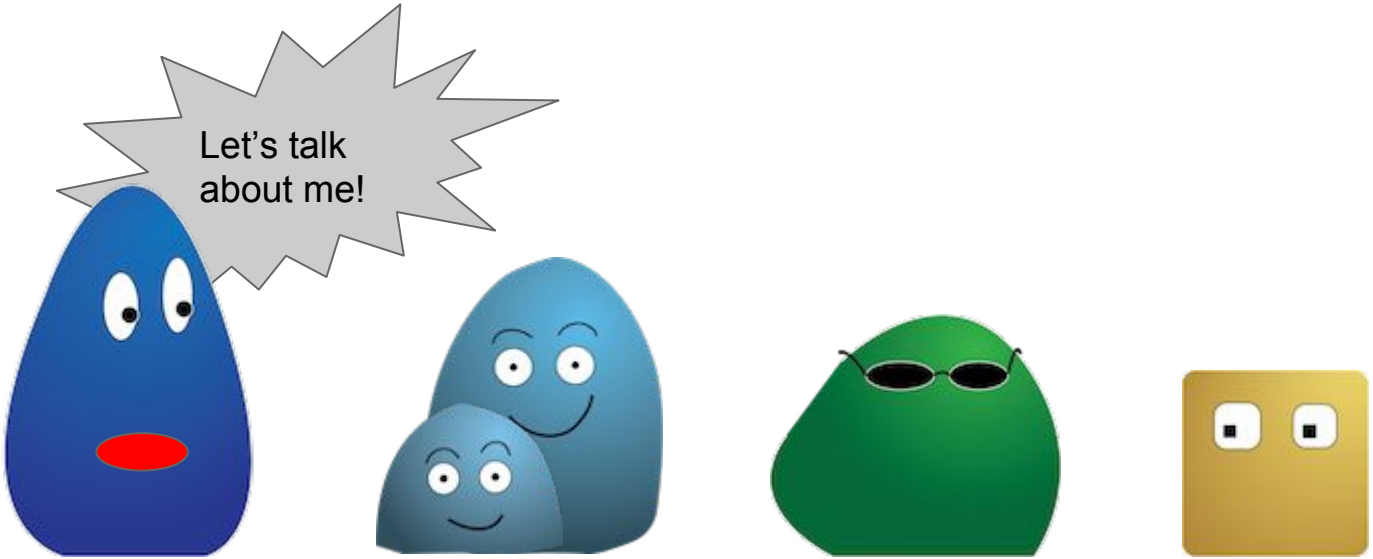
What is a GUID

A **Globally Unique Identifier (GUID)** -

An identifier that follows certain conventions to make it unique within a global context.

In DCPPC, GUIDs are meant to be **globally unique**, **persistent** within a timeframe, and **actionable** on the Web when prefixed by a resolver URI.

Meet the DCPPC GUID cast



DataCite DOI

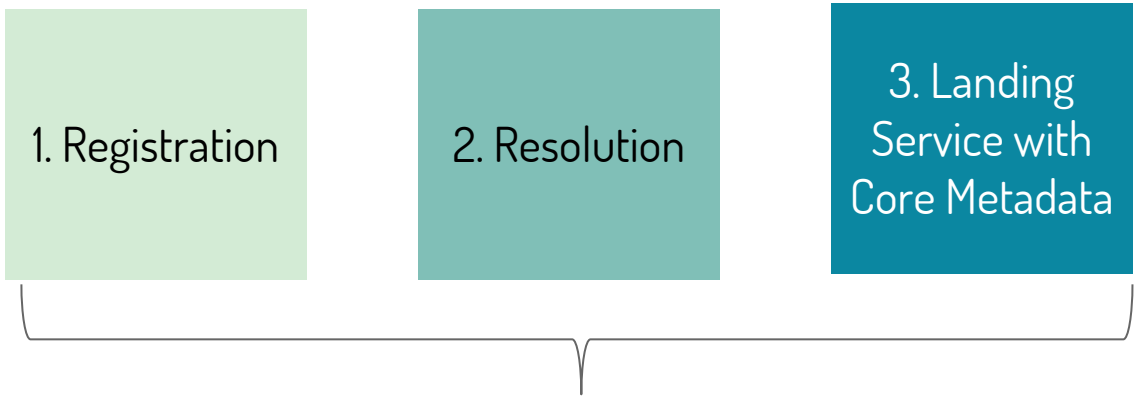
Minid
ARK

Data GUID

Compact
Identifier

A webinar for each GUID type will follow this webinar

What does a GUID need to be DCPPC-compliant



These 3 components enable GUIDs to be **persistent** (within a timeframe) and to **interoperate** across data systems

DOI Registration

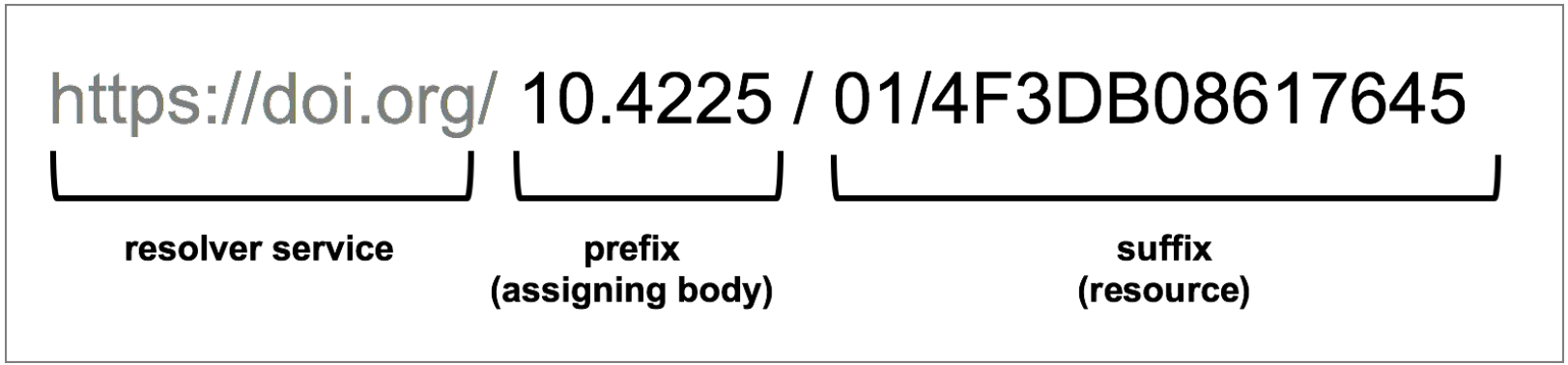
- All DataCite DOIs are registered with metadata conforming to the DataCite metadata schema. Some of these metadata are required.
- Registration of DataCite DOIs can also be done using metadata in schema.org format, which we agreed to use for core metadata in KC2
- 12.8 million DOIs (half of them datasets) have been registered by now 1,600 data centers working with DataCite over the past 10 years.

Core Metadata: required for DOIs

Name	Description	Required
@id	Primary identifier expressed as URL/URI. See discussion of identifier. JSON-LD uses **@id**, RDFa 1.1 uses **resource**, microdata uses **itemid**.	Y
@type	Should in most cases be **Dataset**.	Y
identifier	DOI expressed as URL. One of the identifiers can be a checksum using **propertyValue** with the specific checksum algorithm and checksum.	Y
url	Location of the resource, expressed as HTTP URL. This would normally be the landing page.	Y
includedInDataCatalog	Data provider (data catalog) which hosts this dataset.	Y
Publisher	The publisher of the dataset.	Y
author	The author(s) of the dataset. Schema.org uses **creator** as synonym.	Y
datePublished	Date of first publication.	Y
contentUrl	Actual bytes of the media object, for example the image file or video file.	Y

DOI Syntax and Resolver

A DOI consists of a unique, case-insensitive, alphanumeric character sequence that is divided into three parts separated by a forward slash:

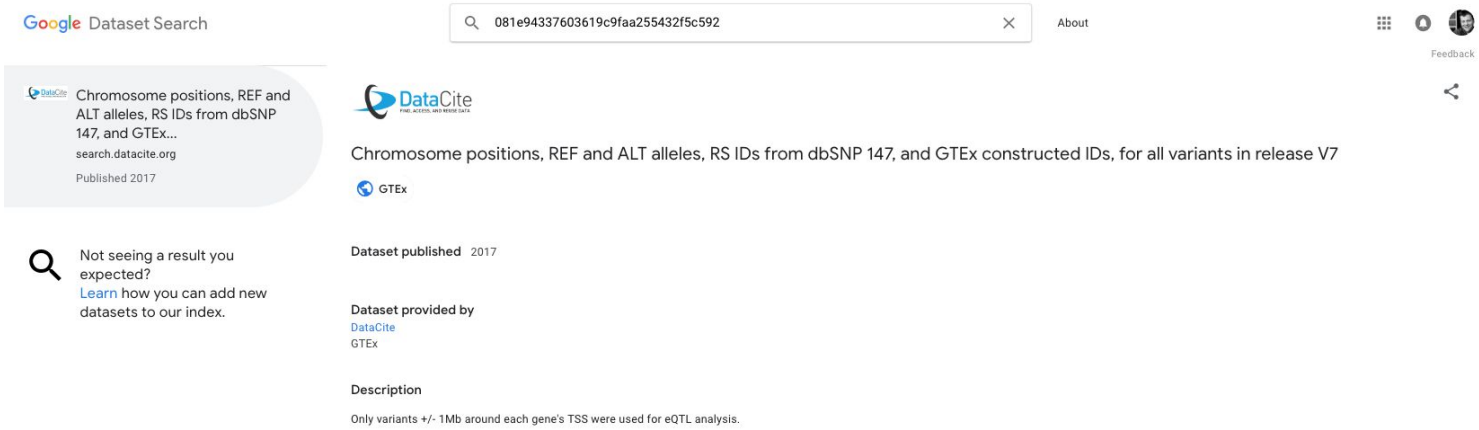


The doi.org resolver is a highly redundant, globally distributed cluster of about 60 servers.

Object Registration Service (ORS)

- **ORS** (Team Sodium) provides registration and landing services for DCPPC GUIDs.
- **KC-compliant** with
  - KC2 (Core Metadata & GUID Services)
  - KC3 (SmartAPIs).
  - KC6 (OAuth2/GlobusAuth).
- **Integrated** with Datacite (DOI) and CDL (Ark) services.

Google Dataset Search



Core metadata in schema.org JSON-LD format and embedded in the ORS landing service are picked up by Google Dataset Search and other indexes.

Use Case: DOIs for GTEx and TOPMed

- We have registered DOIs for public and private datasets provided by data stewards: 26,858 GTEx and 214 TOPMed so far
- GUIDs for MODs will be provided as compact identifiers
- DOIs provide a persistent identifier not specific to a particular full stack. Where available we included other GUIDs (e.g Minid and Data GUID) in the metadata.

DCPPC

11 Na 22.989769 Sodium

Use Case: DOIs for derived datasets

- DOIs should be used for public, persistent datasets that are intended to be included and cited by scholarly articles, and/or linked to contributors (ORCID), funding (Crossref Funder ID), and in the future organizations (ROR) and other entities.