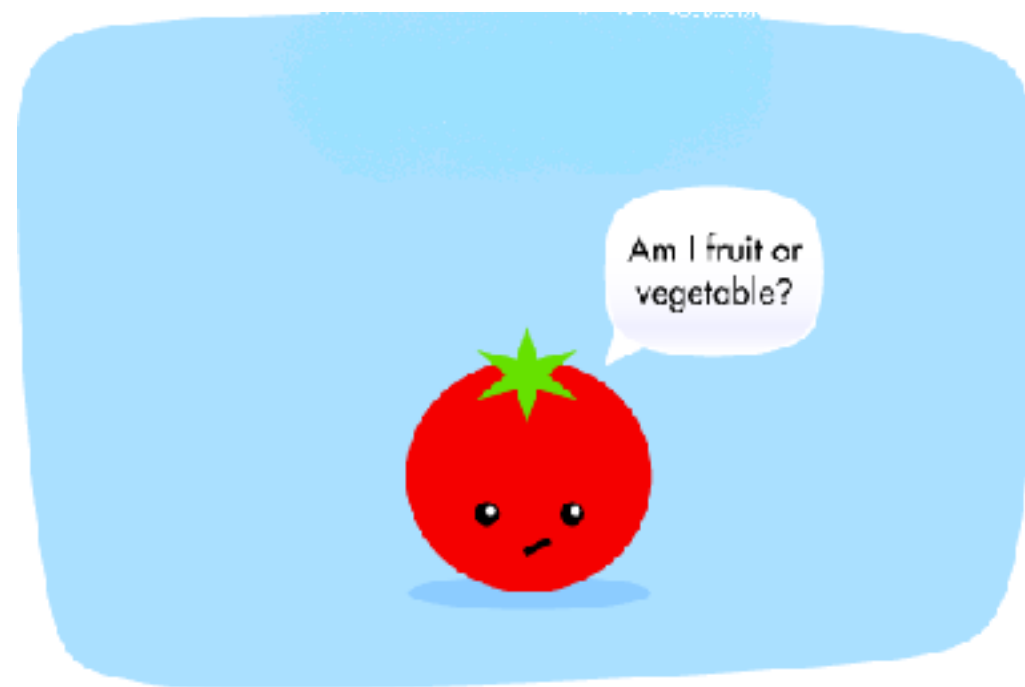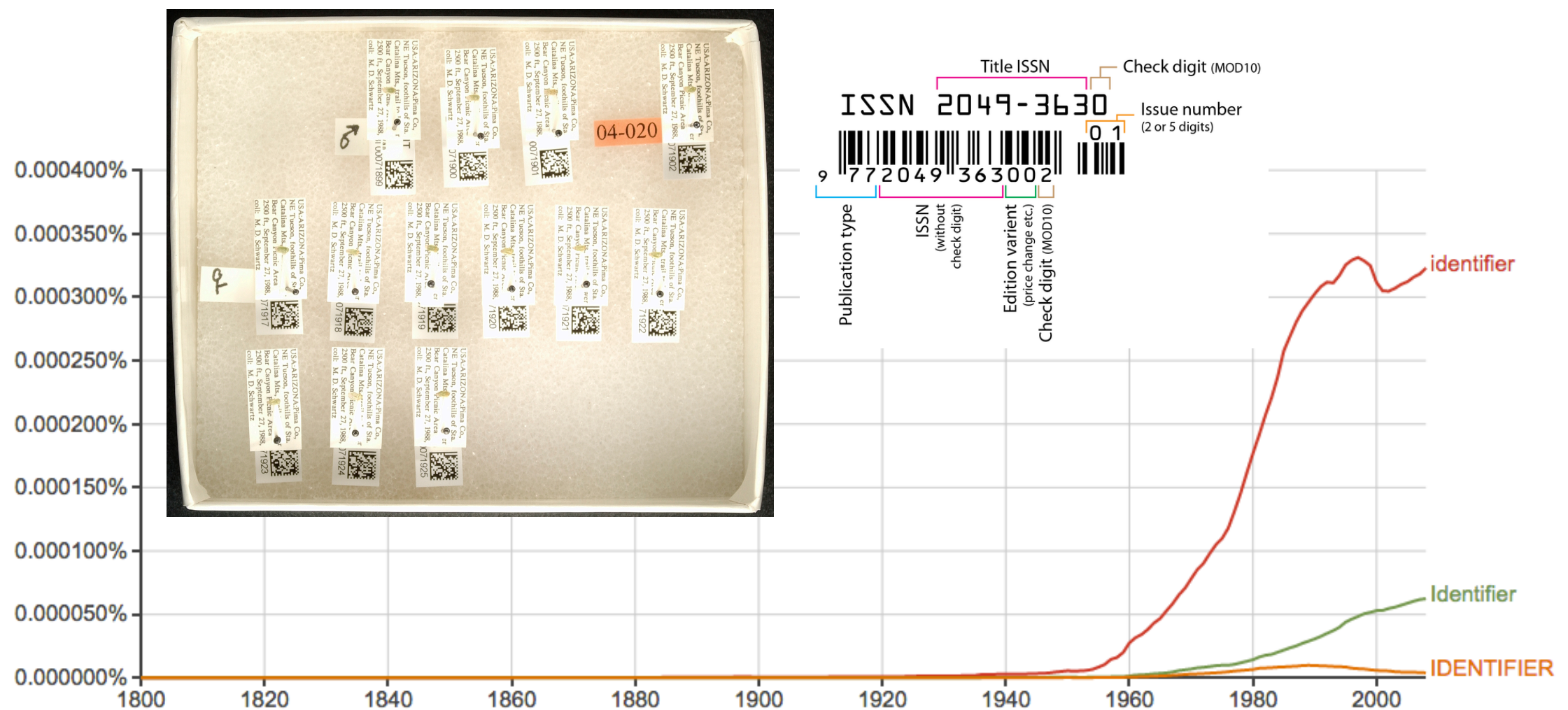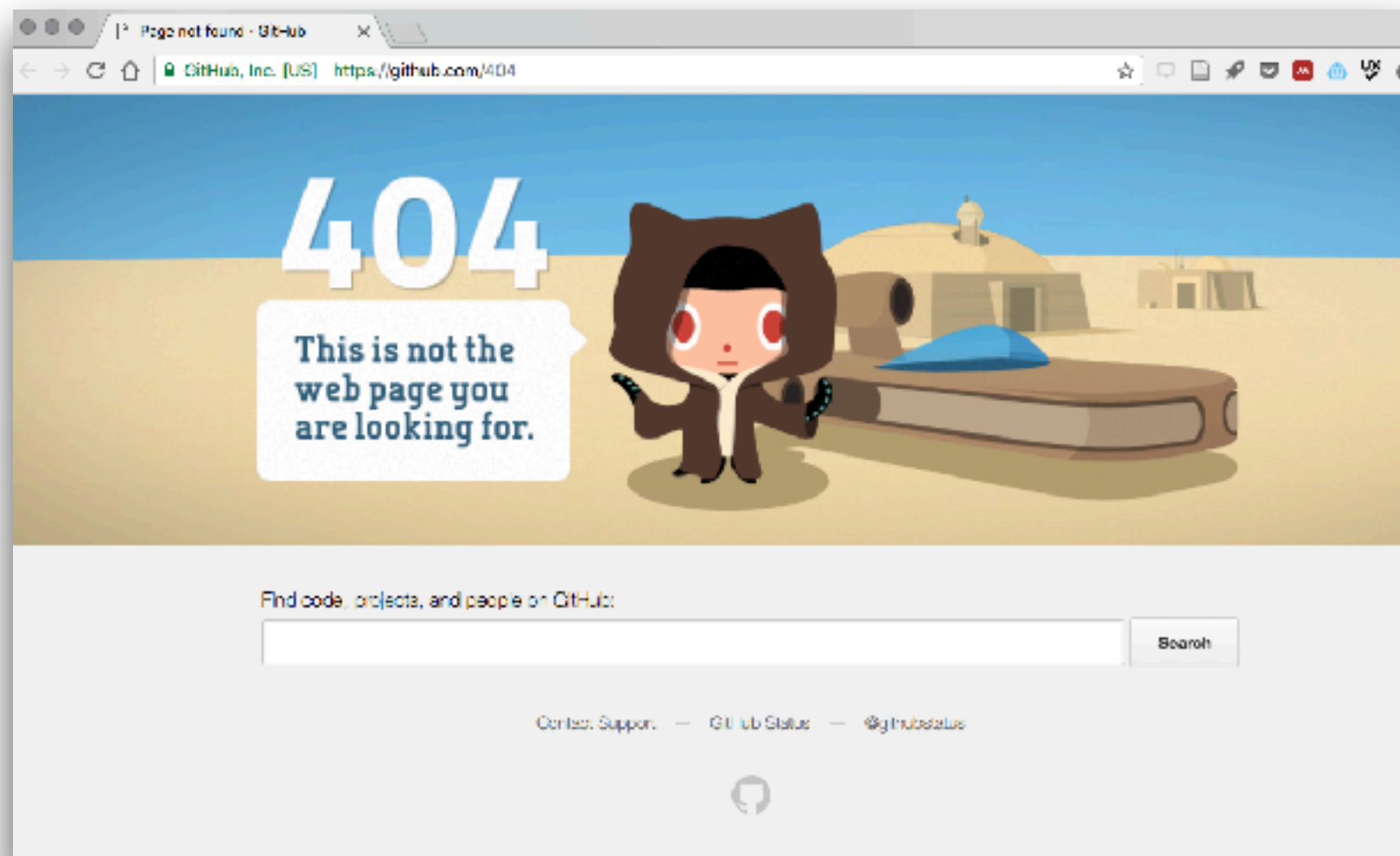# Identifiers + Identity

INFX 551 Wi18

Identifier is "a unique string that identifies either a unique object or a unique class of objects"

Identifiers are everywhere

All URLs are identifiers

- When we point our browser at the wrong identifier, we get undesirable results (404 errors)
- Data are similar to documents on the web- we want to point anyone to the right data with a stable and persistent identifiers

Why IDs matter for data curation?

Test corpus: 3.5 million scholarly articles published 1997-2012

20% had web content that was not accessible for review (e.g. **link rot, content drift**)

If data are playing a role of evidence… we need to make sure that they are stable and accessible for the long-term

Qualities of an Identifier for Digital Data

Unique - universal, non-repeatable string. (No semantics!)

Persistent - immutable at some level of abstraction

Machine Readable - meaning it can be processed and acted upon by software agents

Resolvable - when used with a exchange protocol, the identifier actually goes somewhere.

Identifier Types

Archival Resource Keys (ARKs)

Digital Object Identifiers (DOIs)
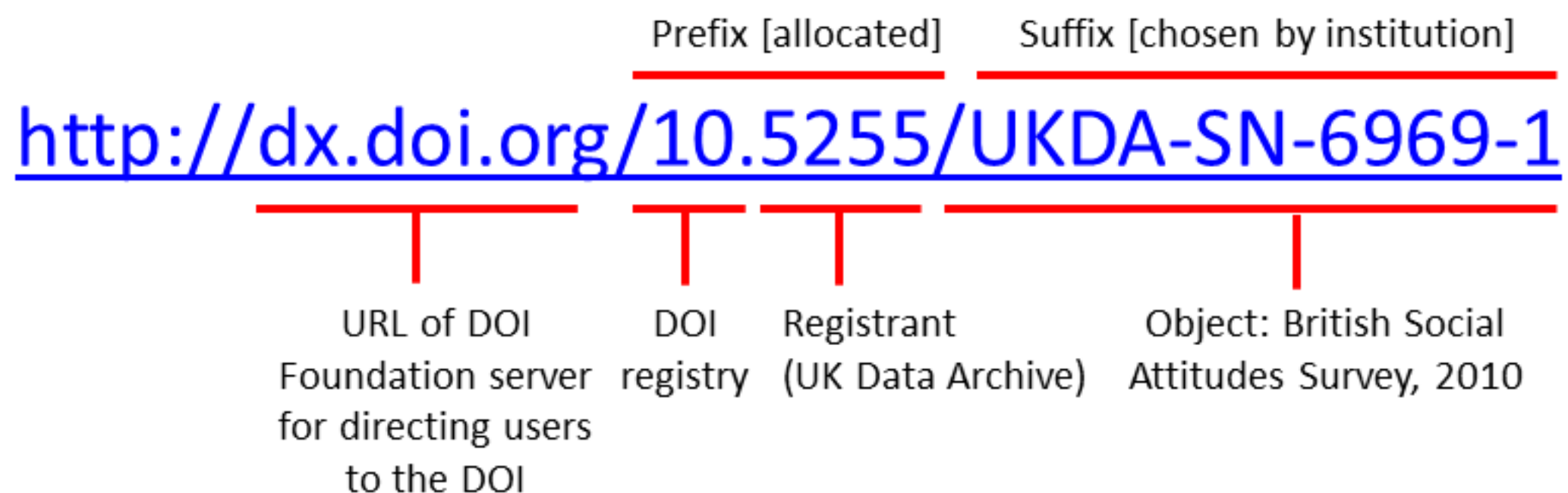
Extensible Resource Identifiers (XRIs)

Handle System (Handles)

Life Science Unique Identifiers (LSIDs)
Object Identifiers (OIDs)

Persistent Uniform Resource Locators (PURLs)

Uniform Resource Identifiers/Names/Locators (URIs/ URNs/URLs)

Universally Unique Identifiers (UUIDs)

Figure 3: Anatomy of a DOI

http://dx.doi.org/10.5255/UKDA-SN-6969-1

Prefix [allocated]

Suffix [chosen by institution]

URL of DOI Foundation server for directing users to the DOI

DOI registry

Registrant (UK Data Archive)

Object: British Social Attitudes Survey, 2010

doi:10.5064/F6H12ZXF

DataCite

EZID
Long-term identifiers made easy
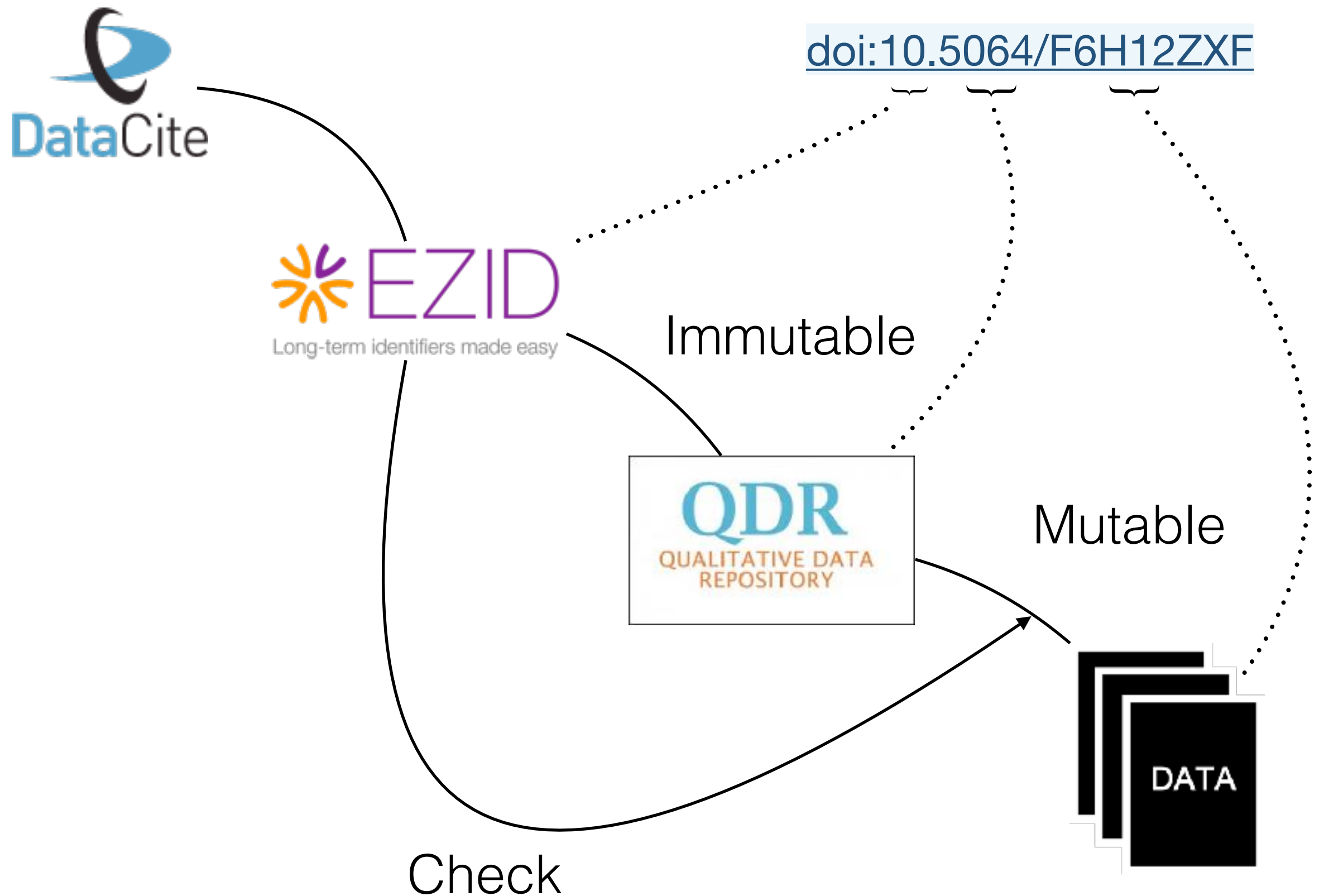
Immutable

QDR
QUALITATIVE DATA
REPOSITORY

Mutable

DATA

Check

ID's are signposts to resources on the web– they need to point to a thing (or give users a clear idea of where to find the thing once they get there)

Just as an address won't prevent a house from being demolished or remodeled, an ID will not prevent "content drift" or data from being deleted

They don't make content more reliable, citable, or trustworthy - they are social cues.