Software source code ID WG

A motivational introduction

Roberto Di Cosmo

roberto@dicosmo.org

December 5th, 2018

Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

How we got in touch with RDA

6th RDA — CNAM, Paris, 09/15

- first contact
- very few CS people, if any!

9th RDA — Barcelona, 04/17

- BoF on Software Preservation …
- 60+ people in the room!

10th RDA — Montreal, 09/17

- Creation of the Software Source Code IG

11th RDA — Berlin, 03/18

- Work on metadata
- Discussion of EU copyright reform

And now…

The *Software Source Code Identification Working Group* (joint RDA/Force 11 effort)

Outline

Systems of identifiers

A *system of identifiers* is

- a set of labels (the identifiers)
- mechanisms to perform :

Generation (minting)	create a new label
Assignment	associate label to object
Retrieval	get object from a label

- optionally, mechanisms to perform:

Verification	check label and object
Reverse Lookup	get label from an object
Description	get metadata of an object

Mechanisms offered in some systems of identifiers

Mech. / System	Handle	DOI	Ark	PURL
Generation	Yes	Yes	Yes	Yes
Assignment	Yes	Yes	Yes	Yes
Retrieval	Yes	Yes	Yes	Yes
Verification	N.A.	N.A.	N.A.	N.A.
Reverse Lookup	N.A.	N.A.	N.A.	N.A.
Description	Yes	Yes	Yes	N.A.

Our challenges in the PID landscape

Typical properties of systems of identifiers

uniqueness, non ambiguity, persistence, abstraction (opacity)

Key needed properties from our use cases

gratis identifiers are free (billions of objects)

integrity the associated object cannot be changed (sw dev, *reproducibility*)

no middle man no central authority is needed (sw dev, *reproducibility*)

we could not find systems with both integrity and no middle man !

An important distinction: DIOs vs. IDOs

The term "Digital Object Identifier" is construed as "digital identifier of an object," rather than "identifier of a digital object" — *Norman Paskin. 2010*

DIO (Digital Identifier of an Object)

digital identifiers for (potentially) non digital objects

- epistemic complexity (manifestations, versions, locations, etc.)
- need an authority to ensure persistence and uniqueness

IDO (Identifier of a Digital Object)

digital identifiers (only) for digital objects

- can provide both integrity and no middle man
- broadly used in modern software development (git, etc.)

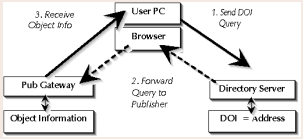
for the core Software Heritage archive, IDOs are enough

Limitations of DIOs

Example: `doi:10.1109/MSR.2015.10`

- to find what 10.1109/MSR.2015.10 is, go to a *resolver* (e.g. doi.org)
- this returns `http://ieeexplore.ieee.org/document/7180064/`
- at this URL we find …

Architecture of the DOI infrastructure



- DOI resolution *can change*
- content at URL *can change*
- no *intrinsic* way of noticing
- persistence based on *good will* of *multiple parties*

Outline

There are many systems of identifiers

- DIOs and IDOs cater to different needs (bit.ly/swhpidpaper)
- IDOs enable integrity and no middle man properties together
  - Software Heritage is using IDOs for billions of objects, today
  - we believe IDOs are appropriate for most digital born content that has a canonical representation

Act before new indicators get standardises (see the OSM!)

- join the RDA SCID WG!
- contribute to understanding how to
  - reference software
  - cite software (not the same!)
- build a consensus