Can Persistent Identifiers be also Cool?

Stefano Bortoli\textsuperscript{1,2}, Paolo Bouquet\textsuperscript{1,2} & Barbara Bazzanella\textsuperscript{1}

\textsuperscript{1} University of Trento (Italy)
\textsuperscript{2} OKKAM srl (Trento, Italy)

8\textsuperscript{th} International Digital Curation Conference (IDCC-2013)
Amsterdam, 14-16 January 2013
Background

OKKAM: Enabling the Web of Entities

*Entity Name System (ENS):*
Managing the lifecycle of identifiers for the semantic web / web of data / linked data

DIGOIDUNA
Study on persistent identifiers for digital objects and authors.

http://www.okkam.org/

http://www.digoiduna.eu/

http://www.alliancepermanentaccess.org/index.php/aparsen/
Positioning

Data Curation

[Persistent] ID mngt

ID for non-digital resources

Long-term positive effects
The context

DIGITAL IDs for DIGITAL and NON-DIGITAL OBJECTS

DECENTRALIZED NETWORKED INFORMATION SYSTEMS, MULTIPLE AUTHORITIES

DIGITAL IDs for DIGITAL and NON-DIGITAL OBJECTS

LOCAL AUTHORITY, SINGLE AUTHORITY INFORMATION SYSTEMS

NON-DIGITAL IDs for NON-DIGITAL OBJECTS

LOCAL AUTHORITY, NON-COMPUTER BASED SYSTEMS

DIGITAL VORTEX
The new fundamental challenge

Digital Identifiers (DIs) are the keys for **cost-effective data management** in digital systems.

We need a solution which can deal with **data & information** created and managed **across**

- **national**
- **organizational**
- **corporate**
- **disciplinary**
- **cultural**
- **technological**...

boundaries (i.e. in networks)

Managing cross-boundary keys to data and resources
Linked Data (Cool URIs)

Built on top of a very robust technical infrastructure (the web)

Based mainly on a **technical approach**: openness, no single point of failure / decentralization, distrust about central authorities

Focused on:
- Data cross linkage & integration / mashup
- Formats
- Shared (semantic) models and vocabularies

Persistent Identifiers (PIIDs)

Built on specialized platforms and systems (e.g. Handle)

Based on **social/organizational principles**: trusted authorities, stakeholders, formal commitments, cost & business models

Focused on:
- Access / preservation / archiving
- Data curation
- Data provenance / quality / authorities
- Value added community services
The OKKAM project (2008-2010): objectives

• Building an **Entity Name System (ENS1.0)** as a sort of DNS for the semantic web:
  – ENS id for an entity $\rightarrow$ IP number for a server
  – Any other web id for the same entity $\rightarrow$ a symbolic name to be mapped to the same ENS id

• Use the ENS as the glue for entities across distributed (heterogeneous) datasets which share common entities (same entity $\rightarrow$ same ID – so no `owl:SameAs` needed, just graph merging!)
• Core ENS services (APIs):
  – ID Lifecycle Management: creation, storage, update, mapping, merge, split, [delete]
  – Entity Matching: given an arbitrary entity description (e.g. a database record), returning the ENS id for the corresponding entity (or more than one in a ranked list)
  – ID mapping: given any entity ID, returning the list of known identifiers for the same entity
• Access Control: open platform, no ownership of identifiers, certificate-based authentication
OKKAM: the ENS1.0

LOCAL Ids OR OKKAM ID

ENS SERVICES (STORAGE, MATCHING, MAPPING, LIFECYCLE)

UNIQUE ID
Example (http://api.okkam.org/search )
Example (http://sig.ma)

ENS id

Alternative IDs from the ENS
The fundamental mistakes of the OKKAM ENS1.0

The ENS1.0 uses resolvable (http) URIs as identifiers and makes data mashup quite straightforward, but it brings in a strong level of centralization (despite its distributed architecture):

- There is no separation between the ID itself and the ID resolver
- Therefore the only resolver for an OKKAM ids is the ENS (and not your local application)

The ENS1.0 offers a technical platform for managing persistent identifiers, but it is not sufficiently trustable:

- Does not have behind it a strong organizational structure
- Its openness does not guarantee data quality
- As a platform, it does not offer value-added services for relevant communities
The ENS 2.0

- Local Resolvers
- HTTP Cool Resolution
- Trusted Authority
- Persistent Unique ID
ENS2.0: Persistent Entity Identifiers

http://www.local-res1.org/peid-8af7c50f-f072-4384-905b-03875c341863

http://www.local-res2.org/peid-8af7c50f-f072-4384-905b-03875c341863

LOCAL RESOLVER 1

LOCAL RESOLVER 2

peid-8af7c50f-f072-4384-905b-03875c341863

PERSISTENT ENTITY IDENTIFIER

http://www.okkam.org/ens/id8af7c50f-f072-4384-905b-03875c341863

DEFAULT RESOLVER
The process for data curators

Value-added community services

Dataset_1

Dataset_2

Dataset_n

Entity alignment (ENS2.0 id)

ENS2.0
Lifecycle management
Entity matching
Resolver registry
Advantages (PIIDs)

• The ENS2.0 ensures the persistence of the binding between an entity and its ID, and only points to data and services about the entity

• The ENS2.0 is a thin, neutral ID management system, no bias towards any vertical application / value added service

• Sustainable cost model, no big costs are involved as it requires limited HW/SW infrastructure and offers only simple core services

• It is going to be managed as an open Public Trust (OKKAM is the Trustee, can be replaced by the Board of Trustees)
Advantages (Cool URIs)

• The ENS2.0 introduces a very light form of centralization (ensure the persistence and uniqueness of the token-entity binding)

• In no sense it can be viewed as an authority for entity data (data stay with local services)

• It links (entities in) data in a way which is much simpler and more maintainable then the standard LOD approach → makes easier to develop value added services from third parties

• It is maintained as an open distributed platform (socially maintained?)
What can we do together?

• We look for partners who can help us to setup the OKKAM Trust and join the Governing Board
• We need institutions & companies who want to experiment with the ENS2.0 for aligning their data with other data (including the Linked Data cloud)
• We look for collaboration with other ID initiatives to ensure the highest level of interoperability
• We need help to improve our entity matching modules, as they are key to the success of the vision
Conclusions

• The fundamental trade-off between centralization / decentralization

• Separation of concerns: ID management should be kept independent from the implementation of value-added community services

• Overcoming organizational barriers: IDs as global keys (semantics), used to implement vertical solutions and services (applications)

• Full compatibility with existing PIDs initiatives (e.g. ORCID for authors), the ENS2.0 as a building block for PIDs interoperability → APARSEN
THANK YOU!

paolo.bouquet@unitn.it
@paolobouquet
#okkam
Identification challenges in the digital era

Managing and preserving new types of widely-distributed, highly volatile, tightly integrated contents

Information access, data sharing, provenance and quality assessment of scientific and non-scientific data across boundaries

Identification of digital and non-digital objects

Need of interlinking contents and related entities like creators, contributors, institutions..

Linking data across repositories and other systems

Datasets access and citability

Reputation and intellectual property
<table>
<thead>
<tr>
<th>Feature</th>
<th>Persistent Identifiers</th>
<th>Cool URIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authority</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Policies</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Level of Trust</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Persistence</td>
<td>YES</td>
<td>NO (?)</td>
</tr>
<tr>
<td>Resolver</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>ID actionability</td>
<td>Partial</td>
<td>YES</td>
</tr>
<tr>
<td>Content change</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Identified entities</td>
<td>Mainly digital objects</td>
<td>Any</td>
</tr>
<tr>
<td>Cross-Linkage</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Content negotiation</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Effort for implementation</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Costs for users</td>
<td>Potentially high</td>
<td>Low</td>
</tr>
</tbody>
</table>