Can Persistent Identifiers be also Cool?

Stefano Bortoli[1,2], **Paolo Bouquet**[1,2] & Barbara Bazzanella[1]

[1] University of Trento (Italy)

[2] OKKAM srl (Trento, Italy)

8th International Digital Curation Conference (IDCC-2013)

Amsterdam, 14-16 January 2013

UNIVERSITY OF TRENTO - Italy

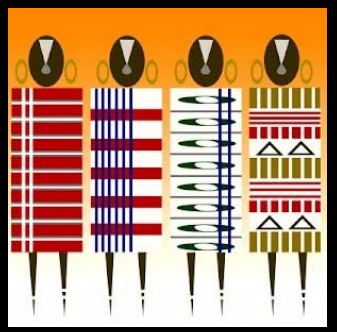
KKAM EMPOWERING THINGS

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

http://www.okkam.org/

DIGOIDUNA
Study on persistent identifiers for digital objects and authors.

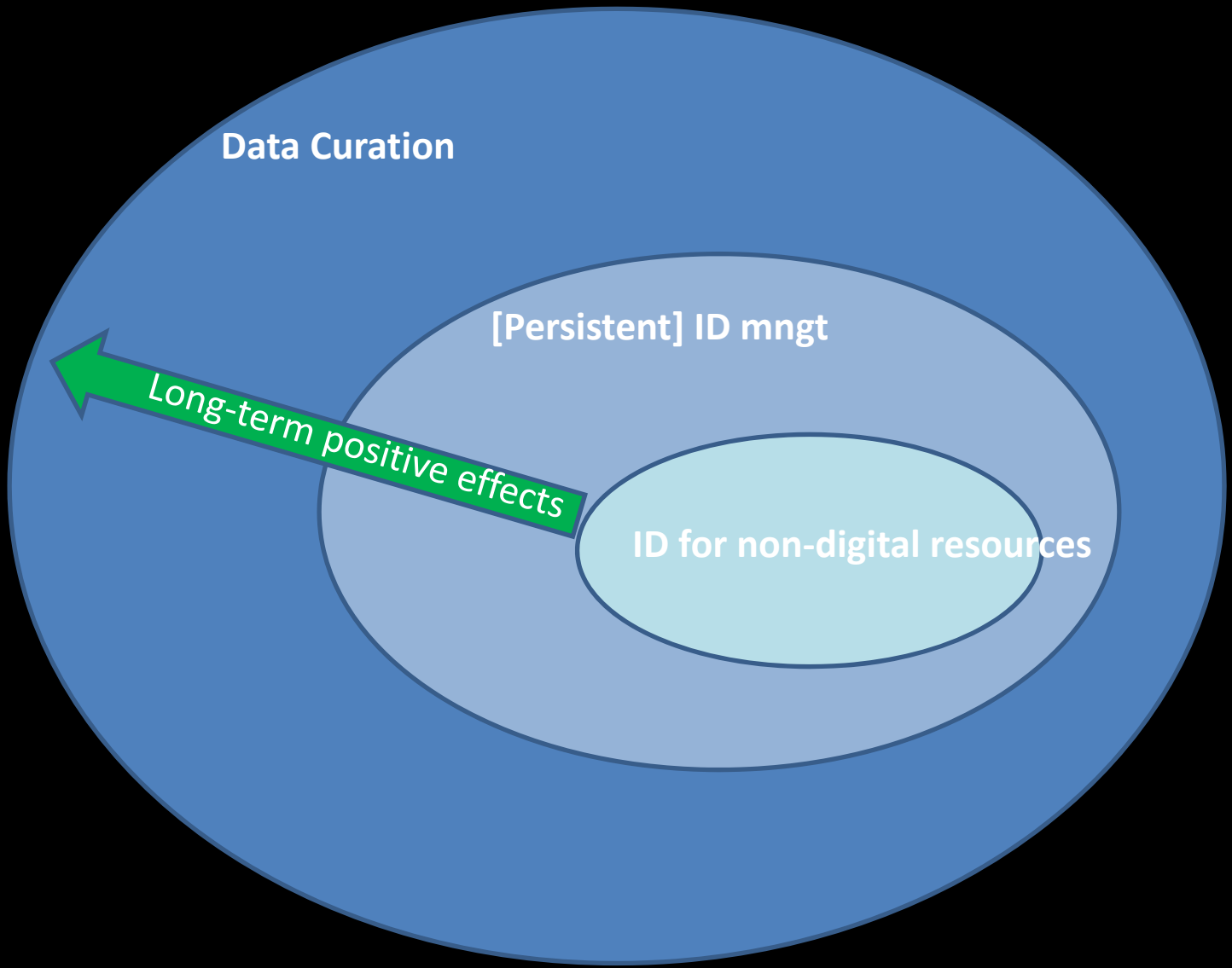


http://www.digoiduna.eu/

Alliance Permanent Access to the Records of Science in Europe Network

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

Positioning



The context

DIGITAL IDs for DIGITAL and NON-DIGITAL OBJECTS,

DECENTRALIZED NETWORKED INFORMATION SYSTEMS, MULTIPLE AUTHORITIES

DIGITAL IDs for DIGITAL and NON-DIGITAL OBJECTS

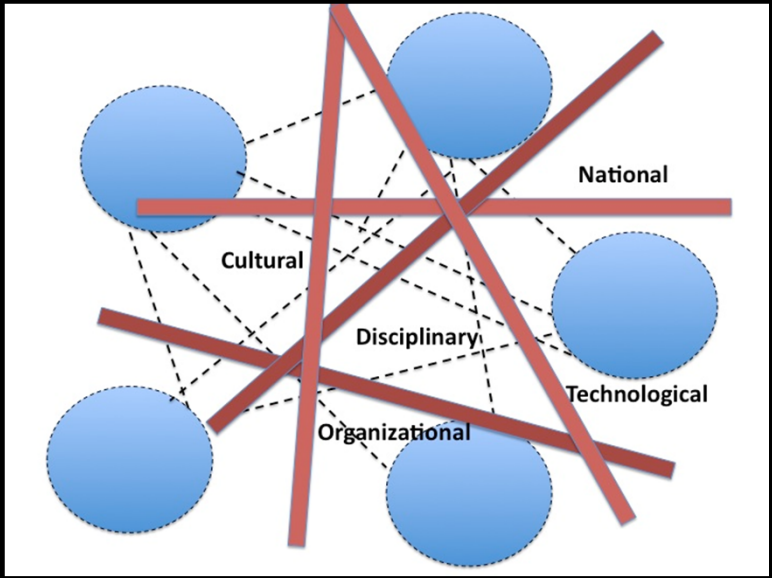
LOCAL , SINGLEAUTHORITY INFORMATION SYSTEMS

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

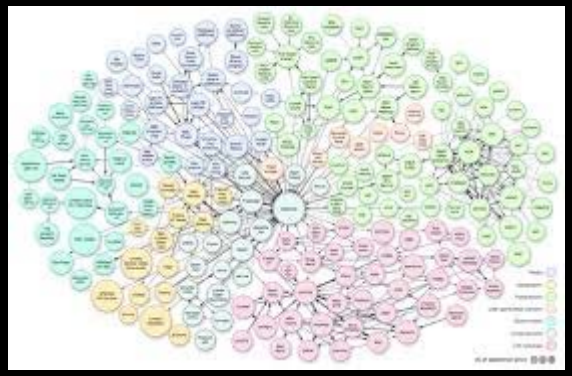
We need a solution which can deal with **data & information** created and managed **across**

- *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 (PIDs)

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

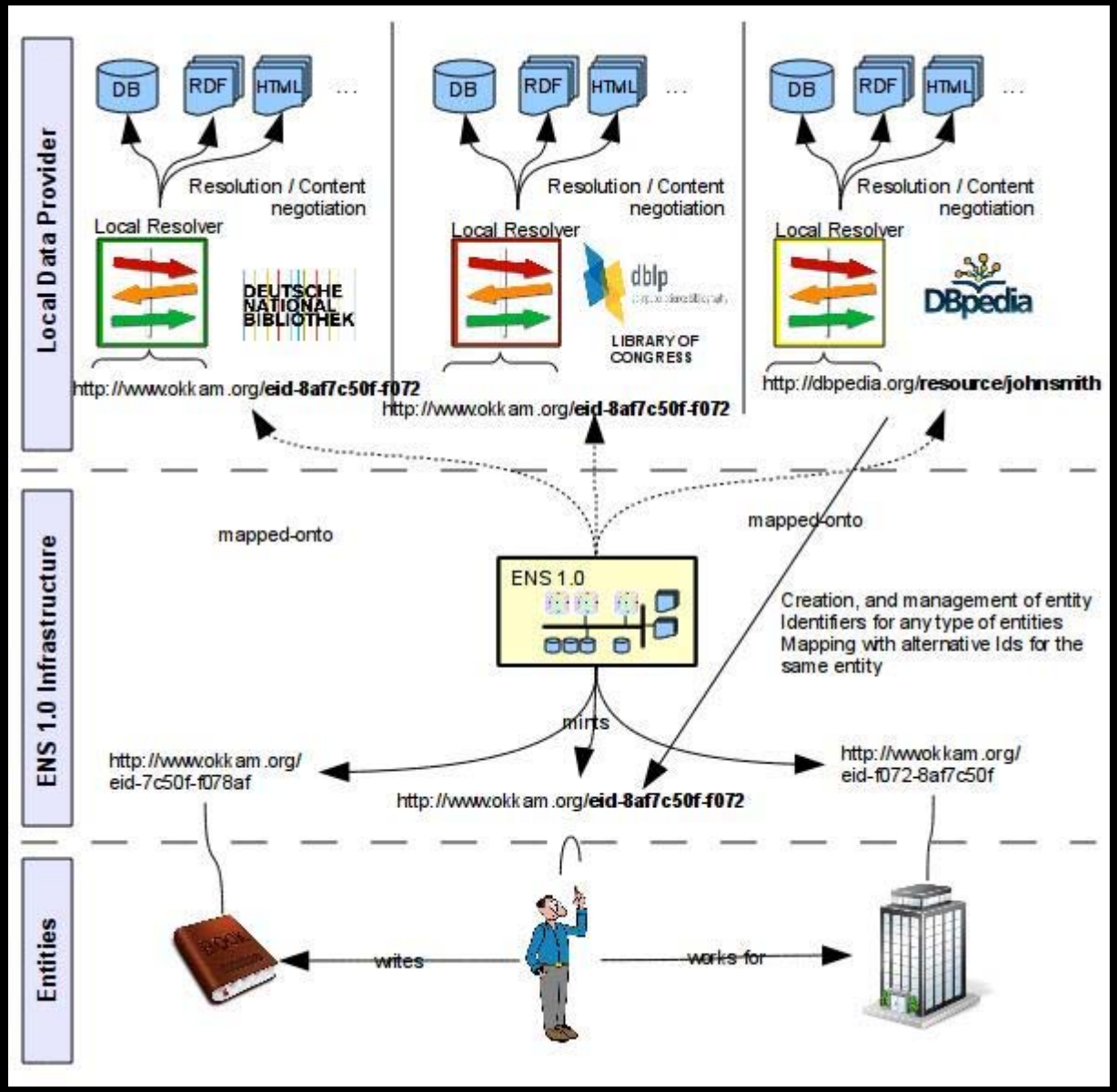
The OKKAM project (2008-2010): objectives

- Building an **Entity Name System (ENS1.0)** as a sort of DNS for the semantic web:
  - ENS id for an entity → IP number for a server
  - Any other web id for the same entity → 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 → 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

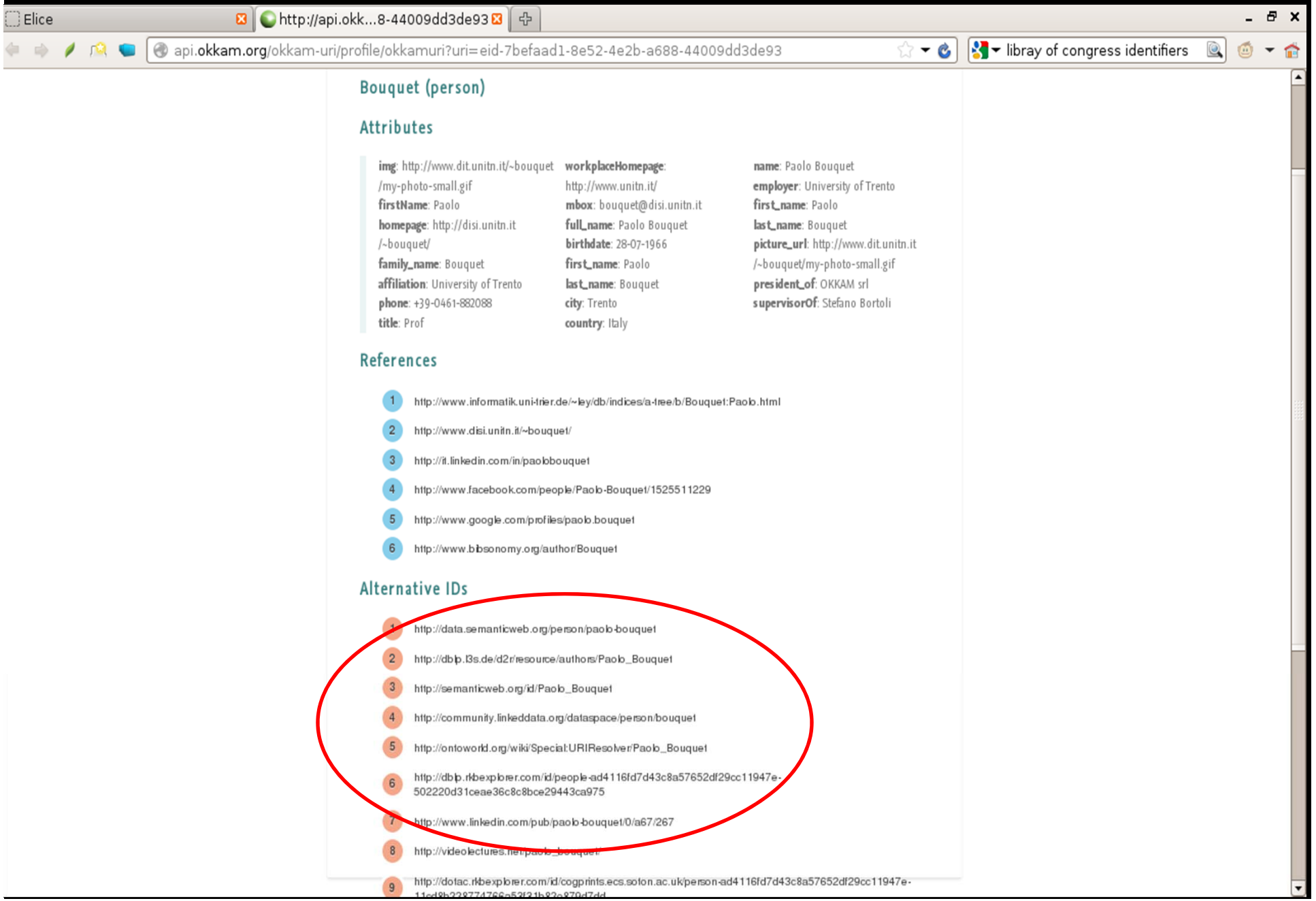


LOCAL Ids OR
OKKAM ID

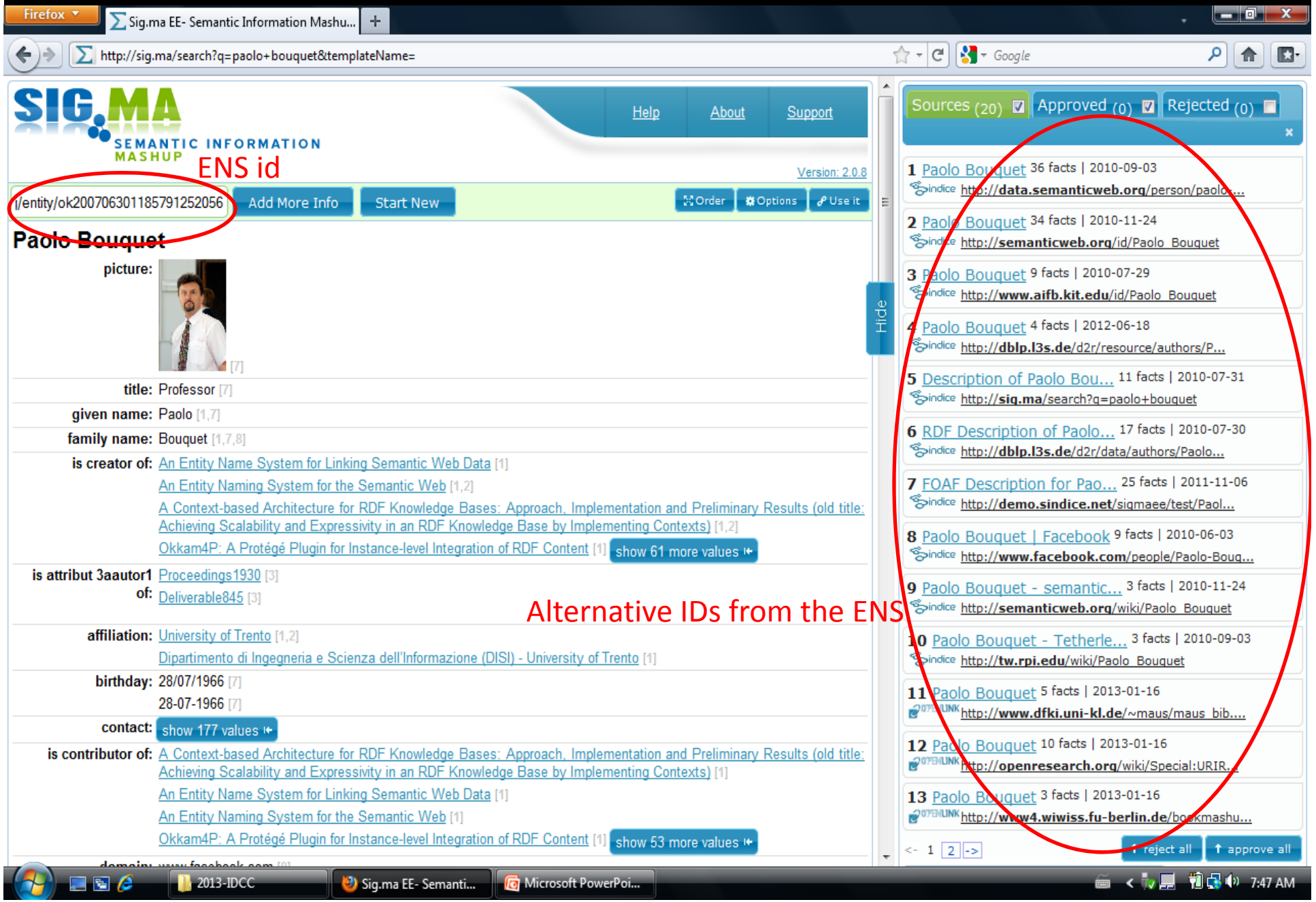
ENS SERVICES
(STORAGE, MATCHING,
MAPPING, LIFECYCLE)

UNIQUE ID

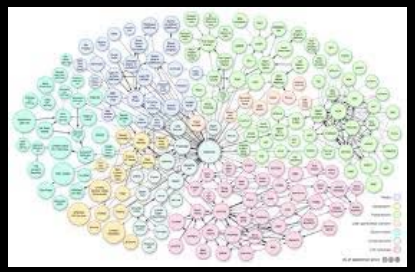
Example (http://api.okkam.org/search)



Example (http://sig.ma)



The fundamental mistakes of the OKKAM ENS1.0



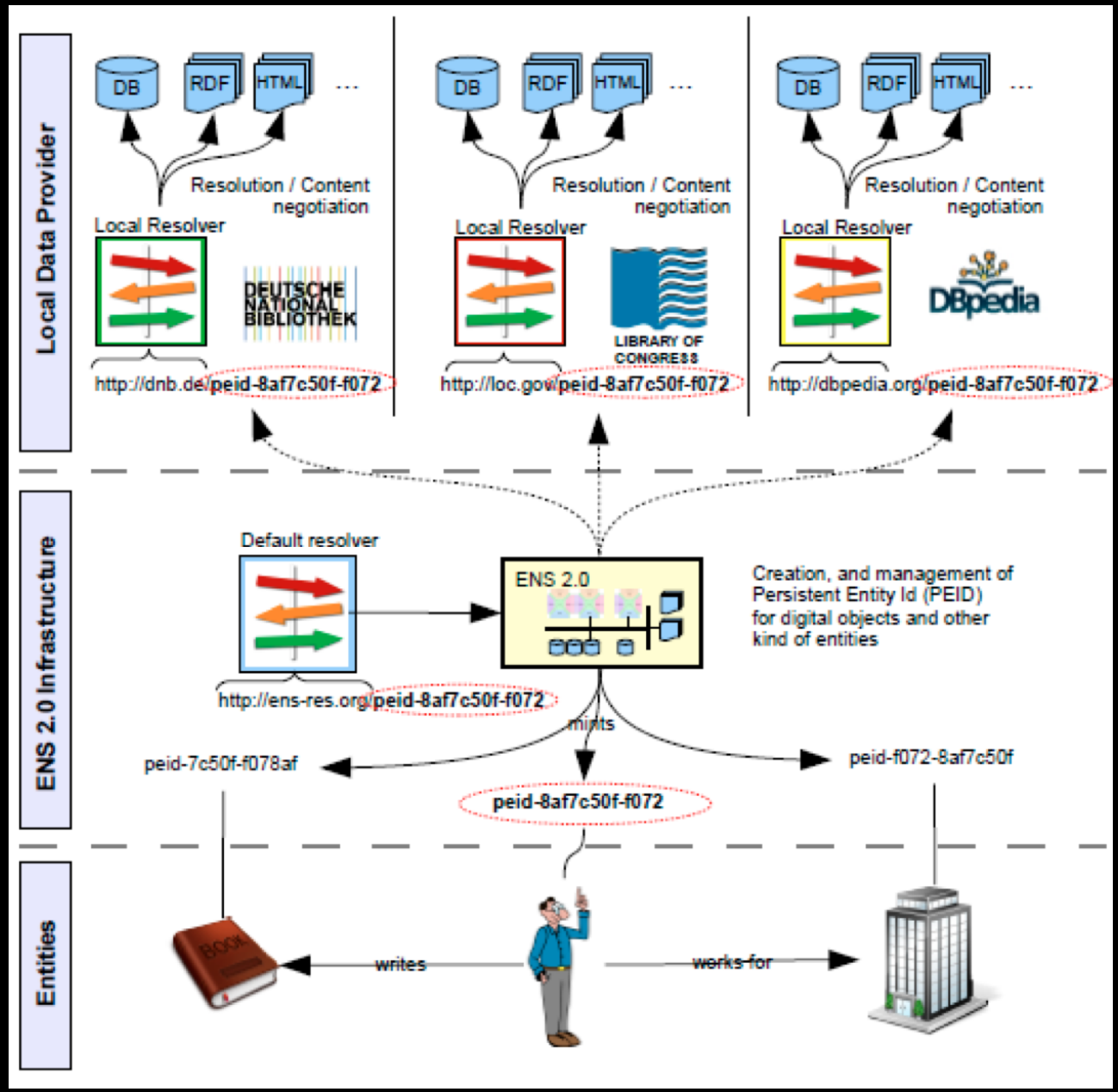
The ENS1.0 uses resolvable (http) URIs as identifiers and makes data mashup quite straightforward, but it brings in a strong level of centralization (despite its distributed architecture):

- 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



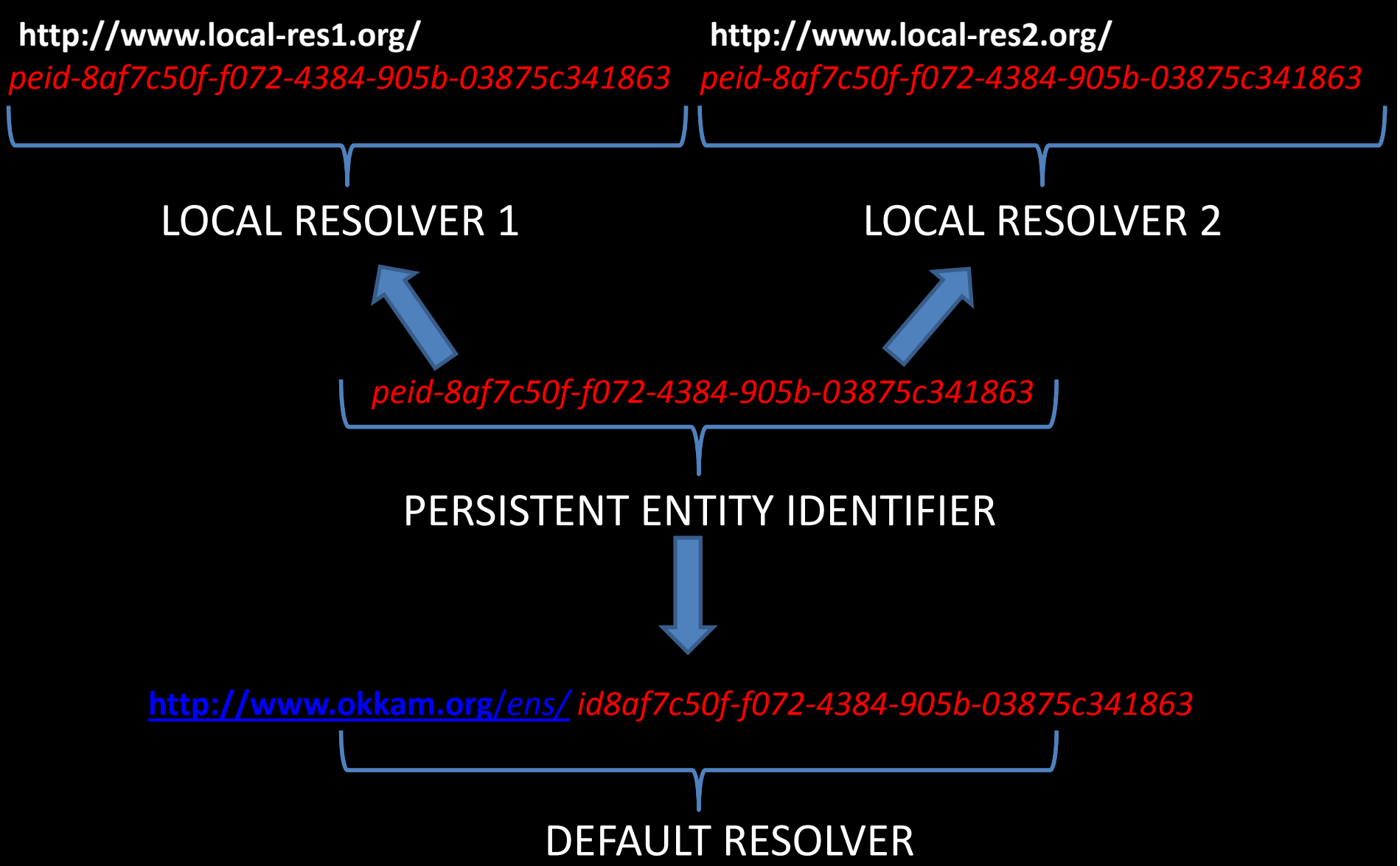
LOCAL RESOLVERS

HTTP COOL RESOLUTION

TRUSTED AUTHORITY

PERSISTENT UNIQUE ID

ENS2.0: Persistent Entity Identifiers



The process for data curators

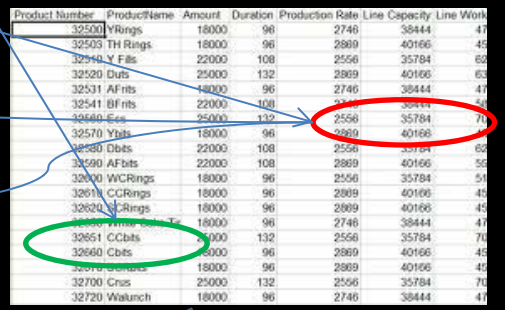
Value-added community services




Dataset_1

Dataset_2

Dataset_n



Entity alignment (ENS2.0 id)

ENS2.0

- Lifecycle management
- Entity matching
- Resolver registry

Advantages (PIDs)

- 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

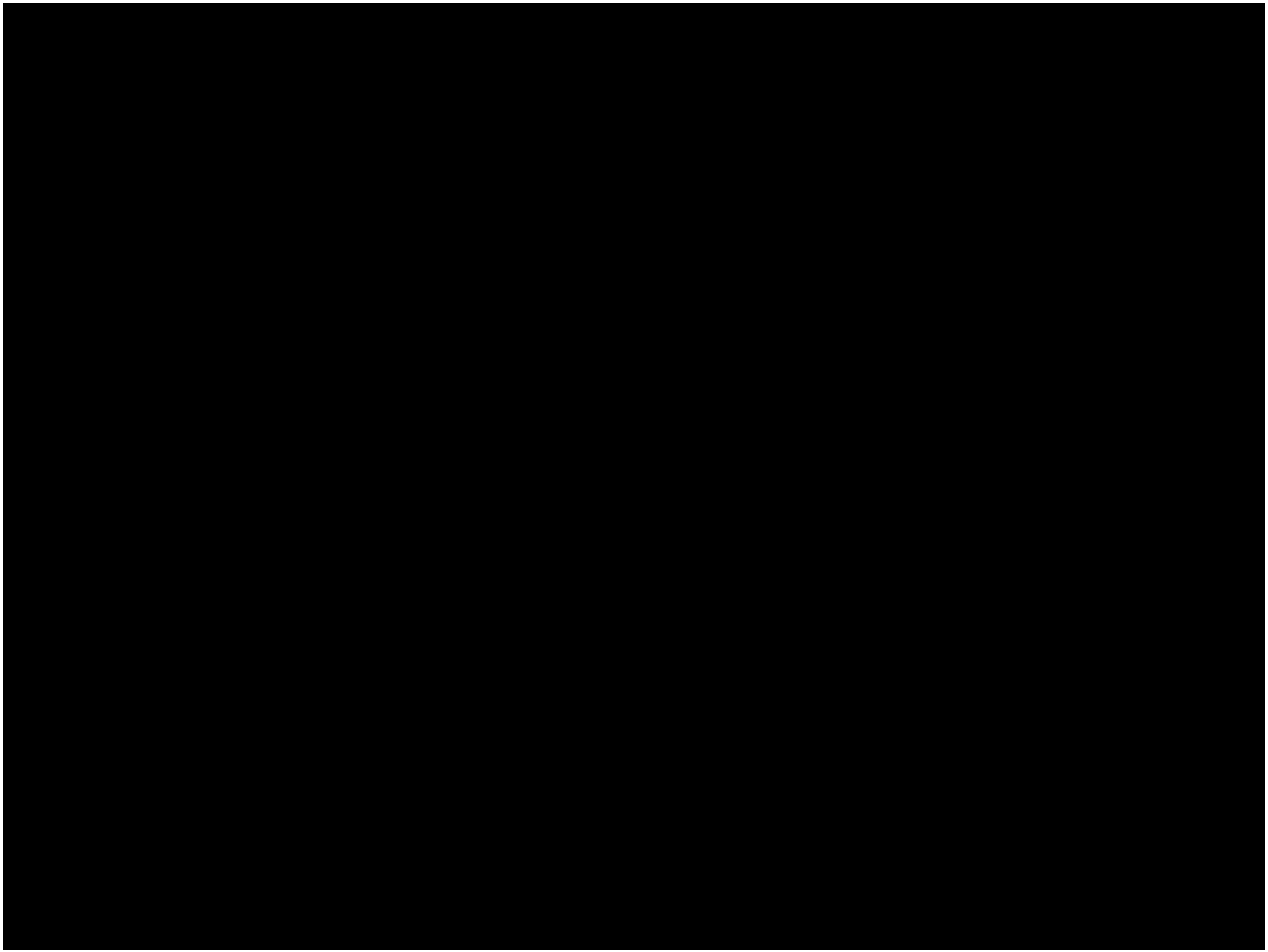
- The fundamental trade-off between centralization / decentralization
- Separation of concerns: ID management should be kept independent from the implementation of value-added community services
- Overcoming organizational barriers: IDs as global keys (semantics), used to implement vertical solutions 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

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

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

Linking data across repositories and other systems

Datasets access and citability

Identification of digital and non-digital objects

Reputation and intellectual property

Comparing Persistent Identifiers and Cool URIs

Feature	Persistent Identifiers	Cool URIs
Authority	YES	NO
Policies	YES	NO
Level of Trust	High	Low
Persistence	YES	NO (?)
Resolver	YES	NO
Uniqueness	YES	NO
ID actionability	Partial	YES
Content change	NO	YES
Identified entities	Mainly digital objects	Any
Cross-Linkage	NO	YES
Content negotiation	NO	YES
Effort for implementation	High	Low
Costs for users	Potentially high	Low