

Can Persistent Identifiers be also Cool?

Stefano Bortoli^{1,2}, **Paolo Bouquet**^{1,2} & Barbara Bazzanella¹

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



Background

OKKAM: Enabling the Web of Entities

Entity Name System (ENS):

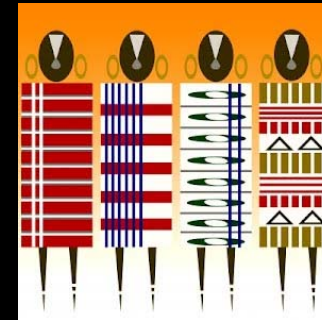
Managing the lifecycle of identifiers for the semantic web / web of data / linked data



<http://www.okkam.org/>

DIGOIDUNA

Study on persistent identifiers for digital objects and authors.

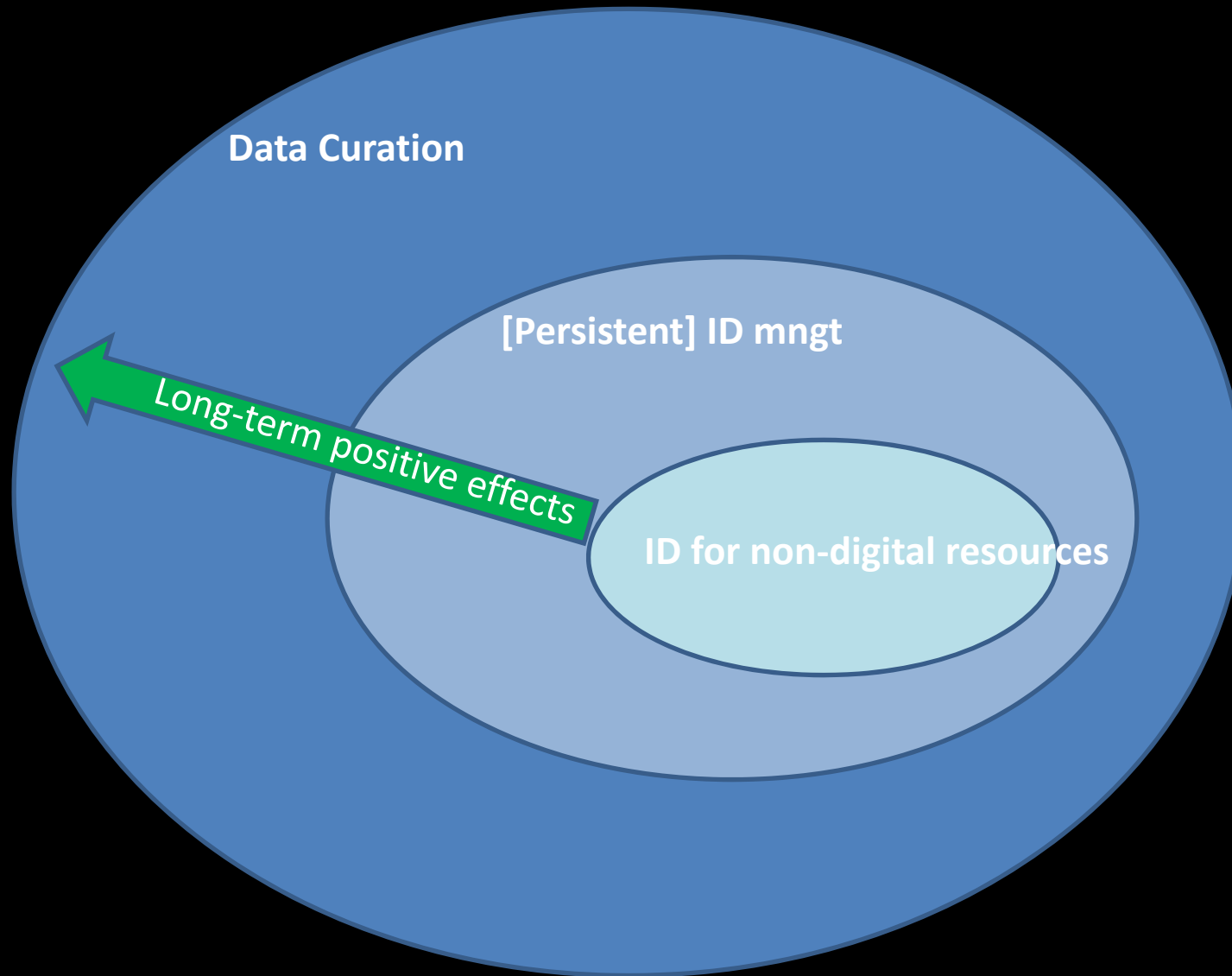


<http://www.digoiduna.eu/>



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

Positioning



The context

DIGITAL IDs for DIGITAL and NON-DIGITAL OBJECTS,
DECENTRALIZED NETWORKED INFORMATION SYSTEMS, MULTIPLE
AUTHORITIES

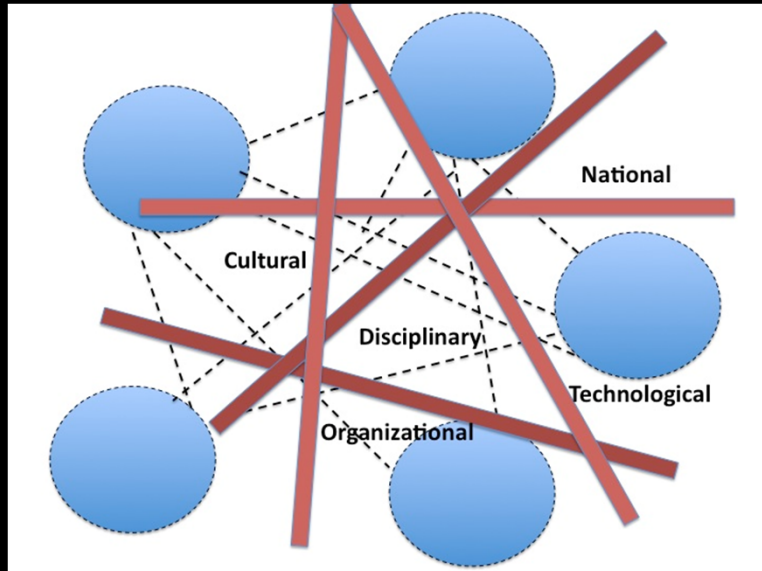
DIGITAL IDs for DIGITAL and NON-DIGITAL OBJECTS
LOCAL, SINGLE AUTHORITY INFORMATION SYSTEMS

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**

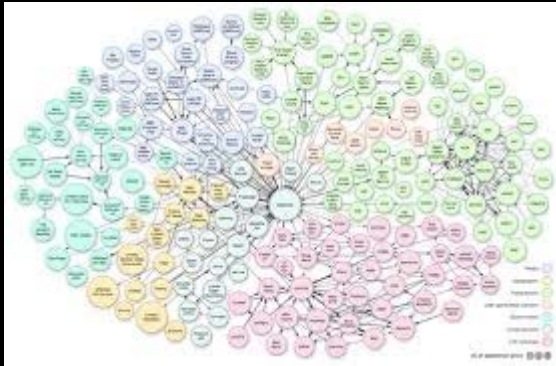
- *national*
- *organizational*
- *corporate*
- *disciplinary*
- *cultural*
- *technological....*

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 (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

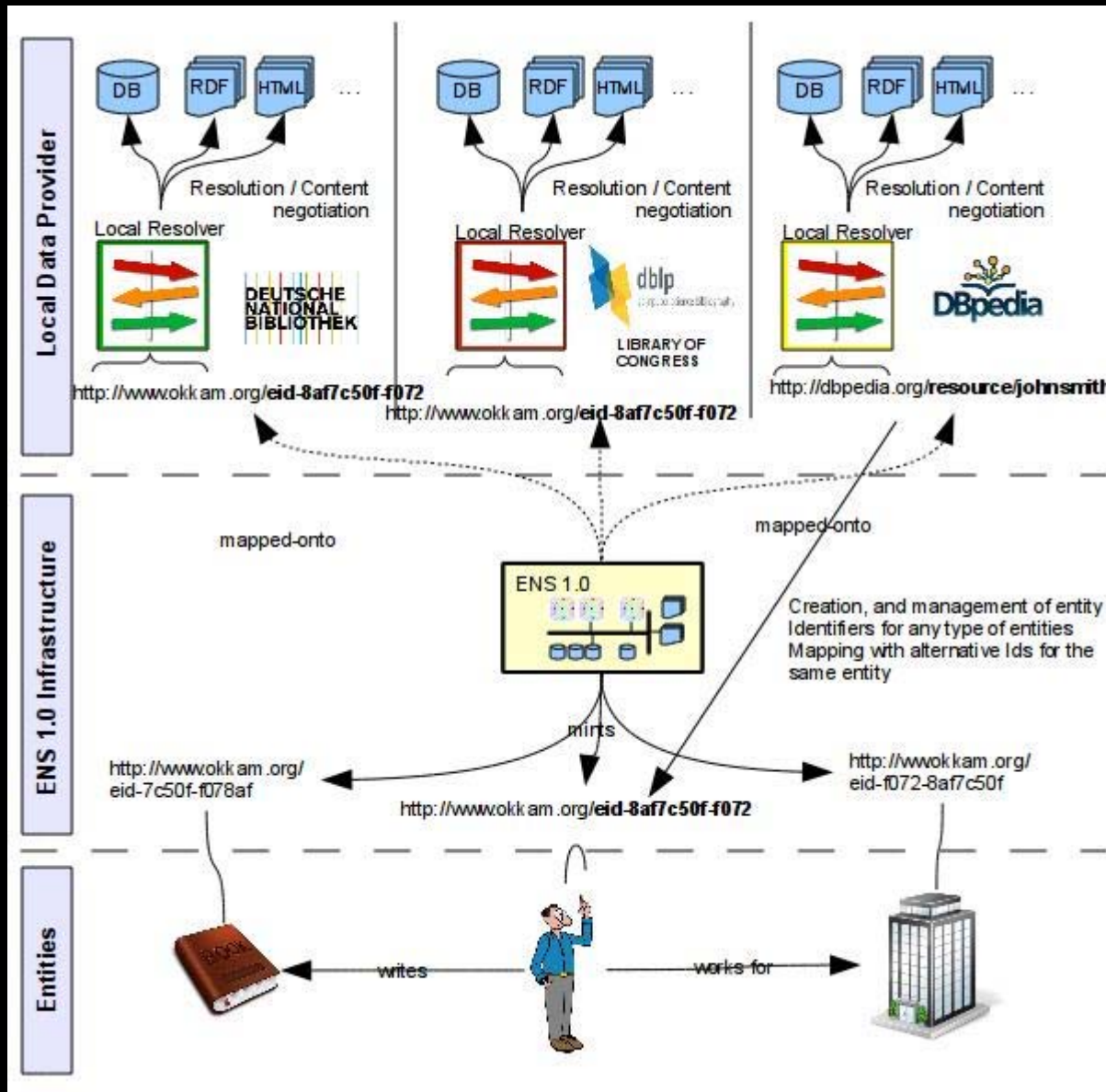
The OKKAM project (2008-2010): objectives

- Building an **Entity Name System (ENS1.0)** as a sort of DNS for the semantic web:
 - ENS id for an entity \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>)

Elice <http://api.okk...8-44009dd3de93>

api.okkam.org/okkam-uri/profile/okkamuri?uri=eid-7befaad1-8e52-4e2b-a688-44009dd3de93 library of congress identifiers

Bouquet (person)

Attributes

img: http://www.dit.unitn.it/~bouquet/my-photo-small.gif	workplaceHomepage: http://www.unitn.it/	name: Paolo Bouquet
firstName: Paolo	mbox: bouquet@disi.unitn.it	employer: University of Trento
homepage: http://disi.unitn.it/~bouquet/	fullName: Paolo Bouquet	firstName: Paolo
family_name: Bouquet	birthdate: 28-07-1966	lastName: Bouquet
affiliation: University of Trento	firstName: Paolo	picture_url: http://www.dit.unitn.it/~bouquet/my-photo-small.gif
phone: +39-0461-882088	lastName: Bouquet	president_of: OKKAM srl
title: Prof	city: Trento	supervisorOf: Stefano Bortoli
	country: Italy	

References

- 1 <http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Bouquet:Paolo.html>
- 2 <http://www.disi.unitn.it/~bouquet/>
- 3 <http://it.linkedin.com/in/paolobouquet>
- 4 <http://www.facebook.com/people/Paolo-Bouquet/1525511229>
- 5 <http://www.google.com/profiles/paolo.bouquet>
- 6 <http://www.bibsonomy.org/author/Bouquet>

Alternative IDs

- 1 <http://data.semanticweb.org/person/paolo-bouquet>
- 2 http://dbpedia.org/resource/Authors/Paolo_Bouquet
- 3 http://semanticweb.org/id/Paolo_Bouquet
- 4 <http://community.linkeddata.org/dataspace/person.bouquet>
- 5 http://ontoworld.org/wiki/Special:URIResolver/Paolo_Bouquet
- 6 <http://dbpedia.org/resource/People-ad4116fd7d43c8a57652df29cc11947e-502220d31ceae36c8c8bce29443ca975>
- 7 <http://www.linkedin.com/pub/paolo-bouquet/0/a67/267>
- 8 http://videolectures.net/paolo_bouquet/
- 9 <http://dotac.rkbexplorer.com/id/cogprints.ecs.soton.ac.uk/person-ad4116fd7d43c8a57652df29cc11947e-11ed8b228774766a52f21b92ca970d7dd>

Example (<http://sig.ma>)


The screenshot shows the Sig.ma Semantic Information Mashup interface. The main content area displays the profile for Paolo Bouquet, with the URL `/entity/ok200706301185791252056` circled in red. The profile includes a picture, title (Professor), given name (Paolo), family name (Bouquet), and various affiliations and contributions. A red circle highlights a list of 13 alternative IDs from the ENS system on the right side of the page.

SIG.MA SEMANTIC INFORMATION MASHUP Version: 2.0.8

ENS id

`/entity/ok200706301185791252056` Add More Info Start New Order Options Use it

Paolo Bouquet

picture: 

title: Professor [7]

given name: Paolo [1,7]

family name: Bouquet [1,7,8]

is creator of: [An Entity Name System for Linking Semantic Web Data](#) [1]
[An Entity Naming System for the Semantic Web](#) [1,2]
[A Context-based Architecture for RDF Knowledge Bases: Approach, Implementation and Preliminary Results \(old title: Achieving Scalability and Expressivity in an RDF Knowledge Base by Implementing Contexts\)](#) [1,2]
[Okkam4P: A Protégé Plugin for Instance-level Integration of RDF Content](#) [1] show 61 more values

is attribut 3aautor1 of: [Proceedings1930](#) [3]
[Deliverable845](#) [3]

affiliation: [University of Trento](#) [1,2]
[Dipartimento di Ingegneria e Scienza dell'Informazione \(DISI\) - University of Trento](#) [1]

birthday: 28/07/1966 [7]
28-07-1966 [7]

contact: show 177 values

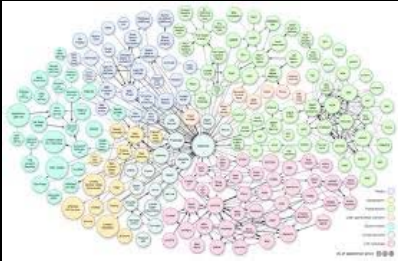
is contributor of: [A Context-based Architecture for RDF Knowledge Bases: Approach, Implementation and Preliminary Results \(old title: Achieving Scalability and Expressivity in an RDF Knowledge Base by Implementing Contexts\)](#) [1]
[An Entity Name System for Linking Semantic Web Data](#) [1]
[An Entity Naming System for the Semantic Web](#) [1]
[Okkam4P: A Protégé Plugin for Instance-level Integration of RDF Content](#) [1] show 53 more values

Alternative IDs from the ENS

- 1 [Paolo Bouquet](#) 36 facts | 2010-09-03
[index](#) <http://data.semanticweb.org/person/paolo...>
- 2 [Paolo Bouquet](#) 34 facts | 2010-11-24
[index](#) http://semanticweb.org/id/Paolo_Bouquet
- 3 [Paolo Bouquet](#) 9 facts | 2010-07-29
[index](#) http://www.aifb.kit.edu/id/Paolo_Bouquet
- 4 [Paolo Bouquet](#) 4 facts | 2012-06-18
[index](#) <http://dblp.l3s.de/d2r/resource/authors/P...>
- 5 [Description of Paolo Bou...](#) 11 facts | 2010-07-31
[index](#) <http://siq.ma/search?q=paolo+bouquet>
- 6 [RDF Description of Paolo...](#) 17 facts | 2010-07-30
[index](#) <http://dblp.l3s.de/d2r/data/authors/Paolo...>
- 7 [FOAF Description for Pao...](#) 25 facts | 2011-11-06
[index](#) <http://demo.sindice.net/siqmaee/test/Paol...>
- 8 [Paolo Bouquet | Facebook](#) 9 facts | 2010-06-03
[index](#) <http://www.facebook.com/people/Paolo-Bouq...>
- 9 [Paolo Bouquet - semantic...](#) 3 facts | 2010-11-24
[index](#) http://semanticweb.org/wiki/Paolo_Bouquet
- 10 [Paolo Bouquet - Tetherle...](#) 3 facts | 2010-09-03
[index](#) http://tw.rpi.edu/wiki/Paolo_Bouquet
- 11 [Paolo Bouquet](#) 5 facts | 2013-01-16
[OPENLINK](#) http://www.dfki.uni-kl.de/~maus/maus_bib...
- 12 [Paolo Bouquet](#) 10 facts | 2013-01-16
[OPENLINK](#) <http://openresearch.org/wiki/Special:URIR...>
- 13 [Paolo Bouquet](#) 3 facts | 2013-01-16
[OPENLINK](#) <http://www4.wiwiss.fu-berlin.de/bookmashu...>

<- 1 2 -> reject all approve all

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):

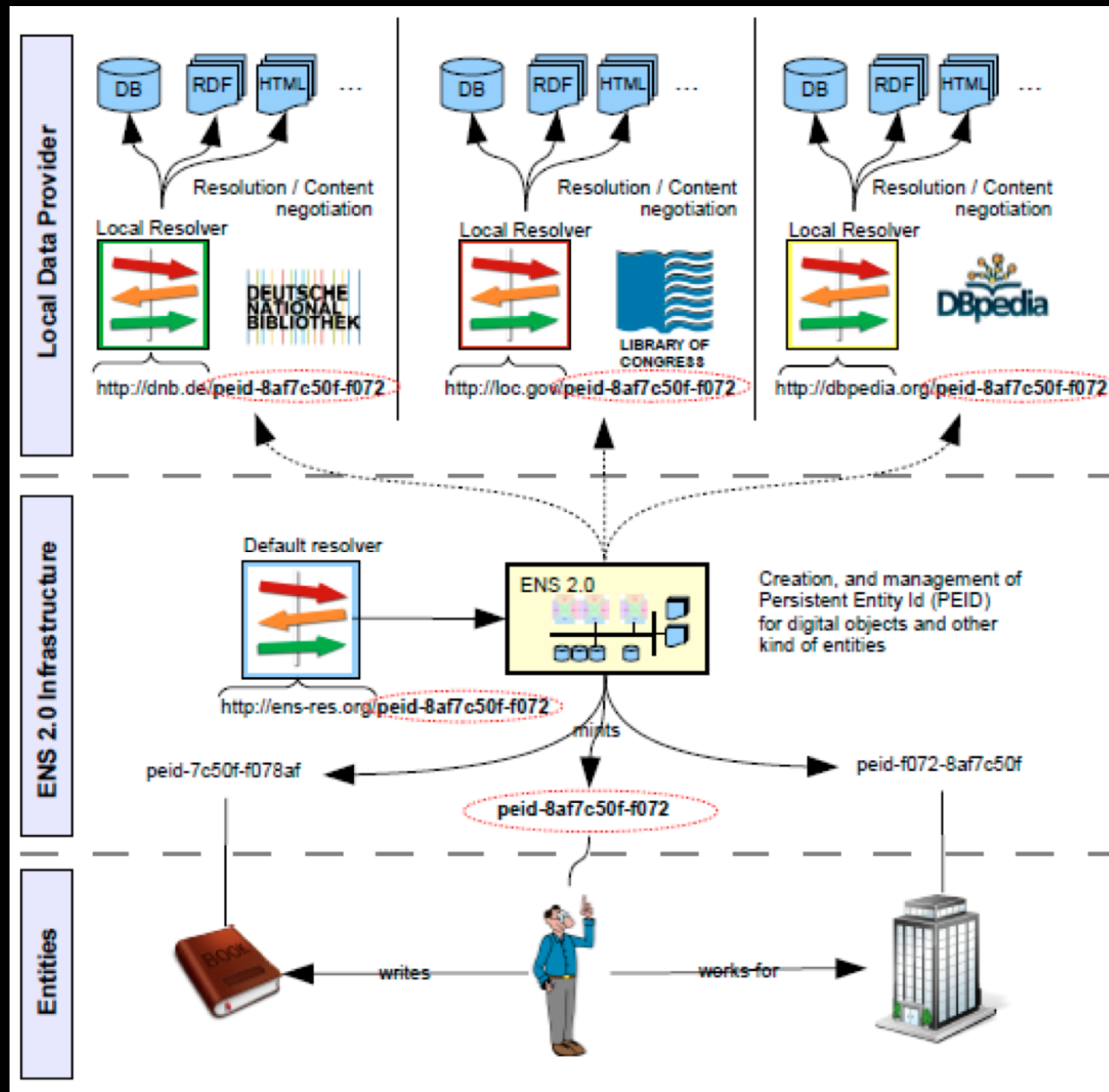
- There is no separation between the ID itself and the ID resolver
- Therefore the only resolver for an OKKAM ids is the ENS (and not your local application)



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

A	B	C	D	E	F	G	H	I	J
12-Jul-1982	27-Aug-2009	2-Mar-2000	10-Nov-2005	4-Jan-1992	23-May-1981				List without duplicates
15-Aug-1989	12-Mar-2001	17-Mar-2001	3-Nov-1956	8-Mar-1978	26-Dec-1984				1982-02-12
24-Feb-1975	25-Mar-2001	6-Dec-2001	26-Sep-1973	2-Jun-1995	2-Oct-1987				1969-09-29
20-Mar-2001	13-Jan-2004	4-Nov-2009	9-Aug-1997	30-Oct-1996	1-Mar-1966				1995-08-22
30-Jul-2006	2-Apr-1985	1-Apr-1999	20-Apr-1975	25-Feb-2000	18-Jan-2068				1979-07-03
20-Oct-1960	10-May-2005	26-Apr-2062	17-May-1982	30-Dec-1988	19-Feb-2007				2000-12-28
22-Feb-1963	5-Jan-1964	11-Mar-1995	5-Aug-2006	3-Jul-2009	1-Sep-2009				1968-03-18
20-Mar-1964	10-May-1985	10-May-2009	10-Sep-2000	27-Jan-1997	26-Dec-1991				1968-10-20
10-Dec-1985	20-Jul-1995	11-Jan-2004	9-Jul-2001	21-Mar-2006	18-Feb-1985				1964-02-28
1-Feb-1990	4-Feb-2001	25-Mar-1968	11-Dec-2000	32-Feb-1999	13-Aug-2009				1995-12-30
16-Aug-2001	2-Apr-2007	25-Jun-2008	9-Aug-2005	13-Aug-2006	15-Aug-2006				1992-02-03
19-Mar-1991	27-Aug-2005	3-Apr-1991	5-Oct-1967	14-Feb-2000	2005-08-14				1995-02-09
26-Mar-1982	11-Jul-2000	5-Sep-2000	2-Sep-2000	31-Jan-1971	2-Jul-2001				1995-03-15
21-Feb-1973	23-Apr-1992	27-Mar-2006	29-Jul-2000	3-Aug-1983	11-Feb-2000				1962-03-24
17	12-Dec-1969	9-Feb-2000	24-Aug-1993	35-May-2004	13-Feb-1981				1975-02-28
20-Aug-1987	5-Oct-2004	33-May-2001	3-Jul-1995	35-May-2004	13-Feb-1981				1967-08-28
3-Aug-2006	2-Apr-1995	14-Dec-2009	23-Jan-2005	23-Jan-2002	6-Nov-1967				1966-06-07
30-Jul-2004	1-Feb-2003	14-May-2007	5-Feb-1968	8-Jul-1976	12-Mar-2009				1985-07-18
28-May-1985	11-Oct-1984	5-Jan-2001	6-Jun-1983	9-Mar-1983	2-May-2007				1985-05-24
11-Jul-1979	22-Jul-2007	25-Jul-2009	11-Oct-1987	20-Jul-1987	11-Jul-2007				1978-02-23

Dataset_2

Name	Owner	Created
MM searchtool_queries_test_csv	vcorn@washington.edu	Mar 26, 2011 9:53 PM
2012 Coding problems	vcorn@washington.edu	Mar 26, 2011 2:56 PM
HIV Surveys - Women's Study	vcorn@washington.edu	Mar 26, 2011 3:40 PM
vcorn bulk load throughput	vcorn@washington.edu	Mar 26, 2011 3:40 PM
hubs throughput statistics	vcorn@washington.edu	Mar 26, 2011 3:40 PM
ribot_ALL_PHYSLUM_COUNT_by_gene_sigs_sigs_read	vcorn@washington.edu	Mar 15, 2011 0:53 PM
ribot_PHYLUM_COUNT	vcorn@washington.edu	Mar 15, 2011 1:18 PM
Triples_3_Biomed_data.csv	vcorn@washington.edu	Mar 15, 2011 2:04 PM
ribot_KOQ_TIGR_GENUS_best_bp_for_each_gene_read	vcorn@washington.edu	Mar 13, 2011 3:12 PM
ribot_ALL_PHYSLUM_COUNT_by_site_sigs_sigs_read	vcorn@washington.edu	Mar 11, 2011 5:57 PM
ribot_ALL_PHYSLUM_COUNT_by_site	vcorn@washington.edu	Mar 11, 2011 5:54 PM
ribot_ALL_PHYSLUM_COUNT	vcorn@washington.edu	Mar 11, 2011 5:13 PM
ribot_ALL_PHYSLUM_COUNT	vcorn@washington.edu	Mar 11, 2011 5:01 PM
ribot_KOQ_PHYSLUM_COUNT	vcorn@washington.edu	Mar 11, 2011 4:47 PM
ribot_KOQ_TIGR_PHYSLUM_best_bp_for_each_gene_read	vcorn@washington.edu	Mar 11, 2011 2:22 PM
ribot_KOQ_PHYSLUM_COUNT	vcorn@washington.edu	Mar 11, 2011 2:18 PM
ribot_KOQ_GENUS_best_bp_for_each_gene_read	vcorn@washington.edu	Mar 11, 2011 1:10 PM

Dataset_n

Product Number	Product Name	Amount	Duration	Production Rate	Line Capacity	Line Work
32500	VRings	18000	96	2746	58444	47
32503	TH Rings	18000	96	2889	40166	45
32560	Y Fils	22000	108	2556	35784	69
32520	Duts	25000	132	2889	40166	63
32531	AFrits	18000	96	2746	38444	41
32541	BFrits	22000	108	2556	38444	51
32646	Ees	25000	132	2556	35784	70
32570	VRings	18000	96	2889	40166	47
32680	Duts	22000	108	2556	38444	62
32690	AFrits	22000	108	2889	40166	50
32700	WCRings	18000	96	2556	35784	51
32610	CCRings	18000	96	2889	40166	45
32620	CCRings	18000	96	2889	40166	45
32630	CCRings	18000	96	2746	38444	47
32650	CCbts	18000	132	2556	35784	70
32651	CCbts	18000	96	2889	40166	45
32700	Crus	25000	132	2556	35784	70
32720	Walunch	18000	96	2746	38444	41

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 than 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

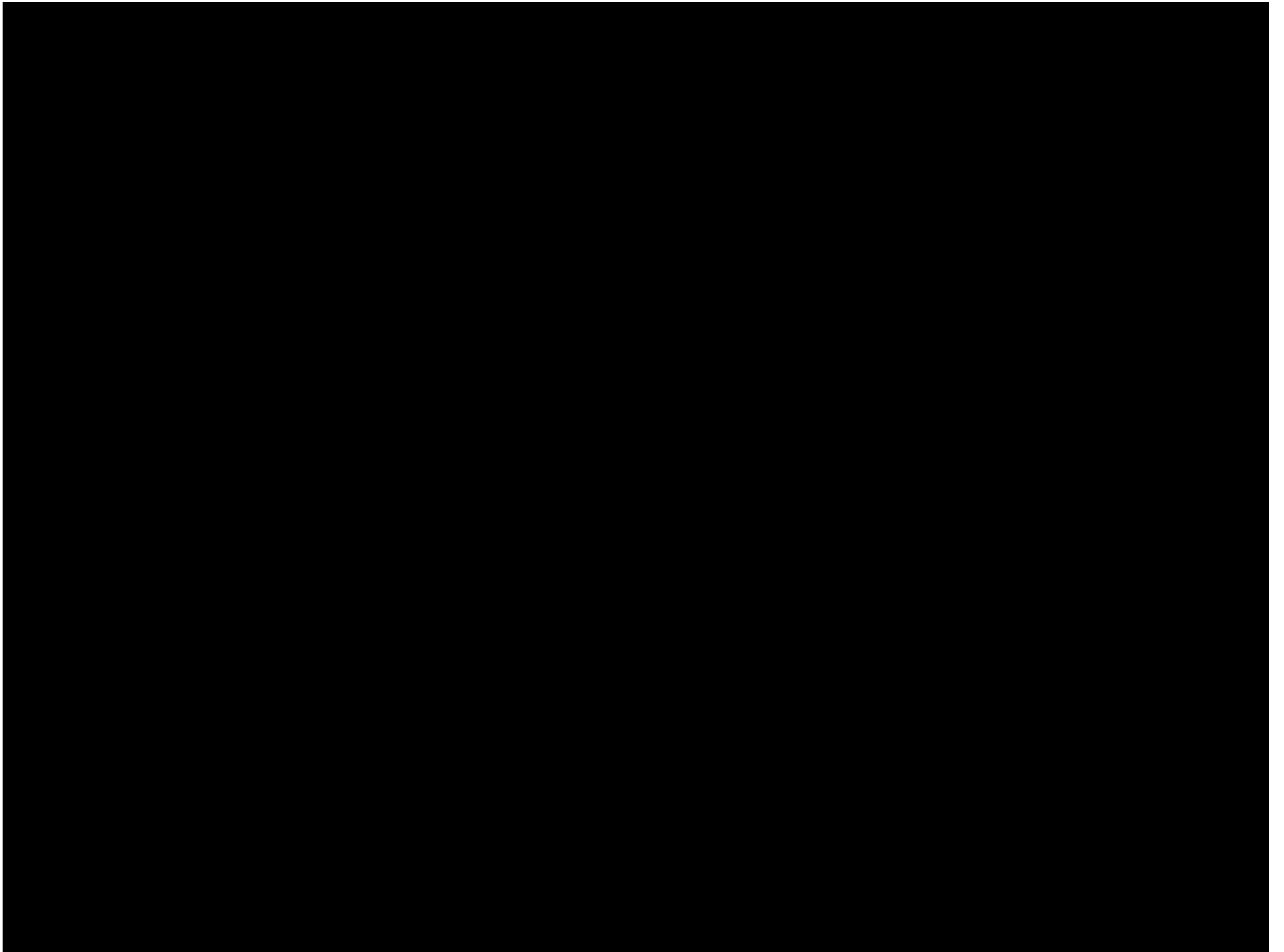
- The fundamental trade-off between centralization / decentralization
- Separation of concerns: ID management should be kept independent from the implementation of value-added community services
- Overcoming organizational barriers: IDs as global keys (semantics), used to implement vertical solutions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

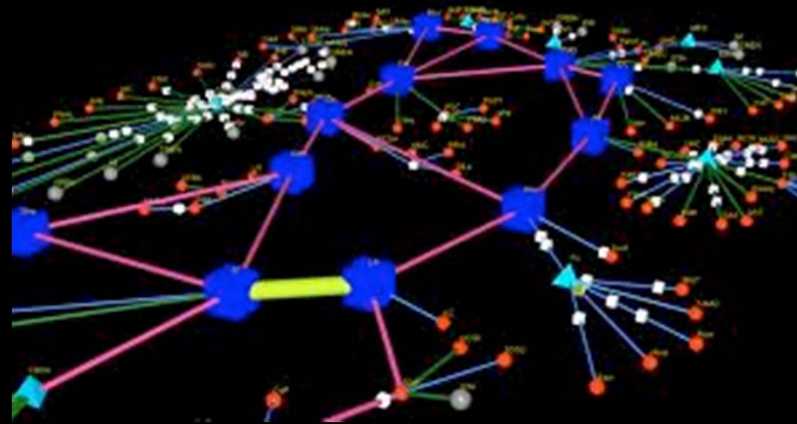


Identification challenges in the digital era

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