



Istituto Centrale
per il Catalogo
e la Documentazione



Convegno

IL PROGETTO ARCO E I DATI APERTI DEL CATALOGO GENERALE DEI BENI CULTURALI

Il Knowledge Graph di ArCo

Andrea Giovanni Nuzzolese

STLab, ISTC-CNR

16 aprile 2019

Istituto centrale per il catalogo e la documentazione

Roma, via di San Michele 18

Sommario

- Produzione di Linked Data
- Scelte di data design
- Entity linking
- ArCo e OntoPiA
- Docker container

Produzione di Linked Data

Obiettivo

- Trasformare le schede del Catalogo Generale dei Beni Culturali in un Knowledge Graph
- Schema di riferimento: moduli della rete di ontologie ArCo
- Paradigma di riferimento: Linked Open Data

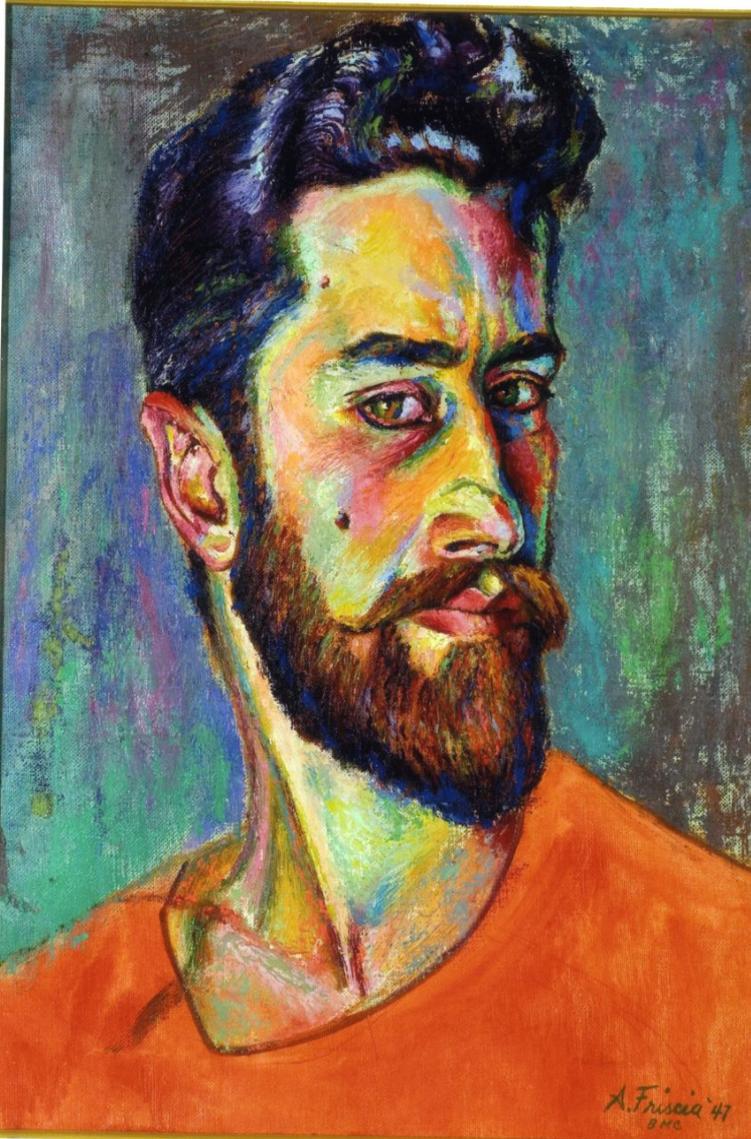
Materiale

- Sorgente dati: ~800 mila di schede catalografiche del Catalogo Generale dei Beni Culturali
- Formato di input: XML

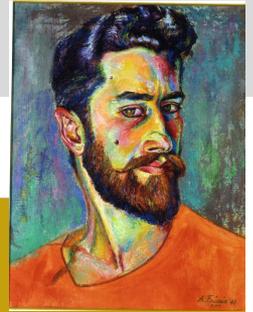
Risultati

- **Ontologie**
 - 395 restrizioni
 - 130 assiomi di disgiunzione
 - 37 allineamenti con 7 ontologie e vocabolari esterni
 - 59 classi e proprietà riusate direttamente da altre ontologie
- **Dati**
 - 173.662.153 triple
 - 24.008 assiomi owl:sameAs per il linking
 - 18.746 entità distinte collegate da altri dataset in LOD

definition «**painting**»
 <OGTD hint="Definizione" />
 title
 <SGTT hint="Titolo" />
 measurements
 unit, height, width
 <MIS hint="MISURE">
 <MISU hint="Unita di misura" />
 <MISA hint="Altezza" />
 <MISL hint="Larghezza" />
 </MIS>
 conservation status «**good**»
 <STCC hint="Stato di conservazione" />
 dating «**20th century**»
 <DTZG hint="Secolo" />
 material, technique
 canvas/ oil-painting
 <MTCI hint="Materiali e tecniche" />
 inscription
 technique «**painting**»
 position «**bottom right**»
 transcript «**A. Friscia '47**»
 <ISR hint="ISCRIZIONE" />
 <ISRS hint="Tecniche" />
 <ISRP hint="Posizione" />
 <ISRI hint="Trascrizione" />



unique code, name «**Albert Friscia**»
 historical data «**active from 1935 to 1989**»
 <AUTORE hint="Autore" />
 <NCUN hint="Codice univoco ICCD" />00001499</NCUN>
 <AUTN hint="Nome scelto" />Albert Friscia</AUTN>
 <AUTA hint="Dati anagrafici" />Attivo dal 1935 al 1989</AUTA>
 acquisition type «**donation**», previous owner «**Di Bello, Lidia**»
 location
 <ACQUISIZIONE hint="Acquisizione" />
 <ACQT hint="Tipo acquisizione" />donazione</ACQT>
 <ACQN hint="Nome" />Di Bello, Lidia</ACQN>
 <ACQD hint="Data acquisizione" />10/ 02/ 2000</ACQD>
 <ACQL hint="Luogo acquisizione" />RM/Roma/Via Muratte,94</ACQL>
 previous location type «**previous location**»
 <PRCT hint="Tipologia" />private house</PRCT>
 <PRCD hint="Denominazione" />house of Mr and Mrs Friscia</PRCD>
 <PRCU hint="Denominazione via" />Via Muratte,94 Roma</PRCU>
 <PRCS hint="Specifiche" />inside, 2° floor</PRCS>
 photographic documentation
 category «**attached**», type «**color**»
 owner agency
 <DOCUMENTAZIONE FOTOGRAFICA hint="Documentazione fotografica" />
 <FTAX hint="Genere" />documentazione allegata</FTAX>
 <FTAP hint="Tipo" />fotografia colore</FTAP>
 <FTAE hint="Ente proprietario" />SPSAE MT</FTAE>



Self portrait, Autoritratto (dipinto)

<https://w3id.org/arco/resource/HistoricOrArtisticProperty/1700168615>

ENTITÀ DI TIPO: **HistoricOrArtisticProperty**

rdfs:comment

dipinto

rdfs:label

Self portrait, Autoritratto (dipinto)

arco:catalogueNumber

00168615

arco:regionIdentifier

17

arco:uniqueIdentifier

1700168615

a-cd:subject

Autoritratto

a-cd:title

Self portrait

arco:description

La tela raffigura l'autoritratto del pittore, dipinto a mezzo busto, con il capo ruotato di tre quarti verso sinistra. Il colore arancio acceso della maglietta spicca sul fondo bluastrò della tela; il volto è segnato da tratti spigolosi, da una folta barba e da un neo sullo zigomo destro.

rdf:type

arco:HistoricOrArtisticProperty

↳ Bene Storico o Artistico

—

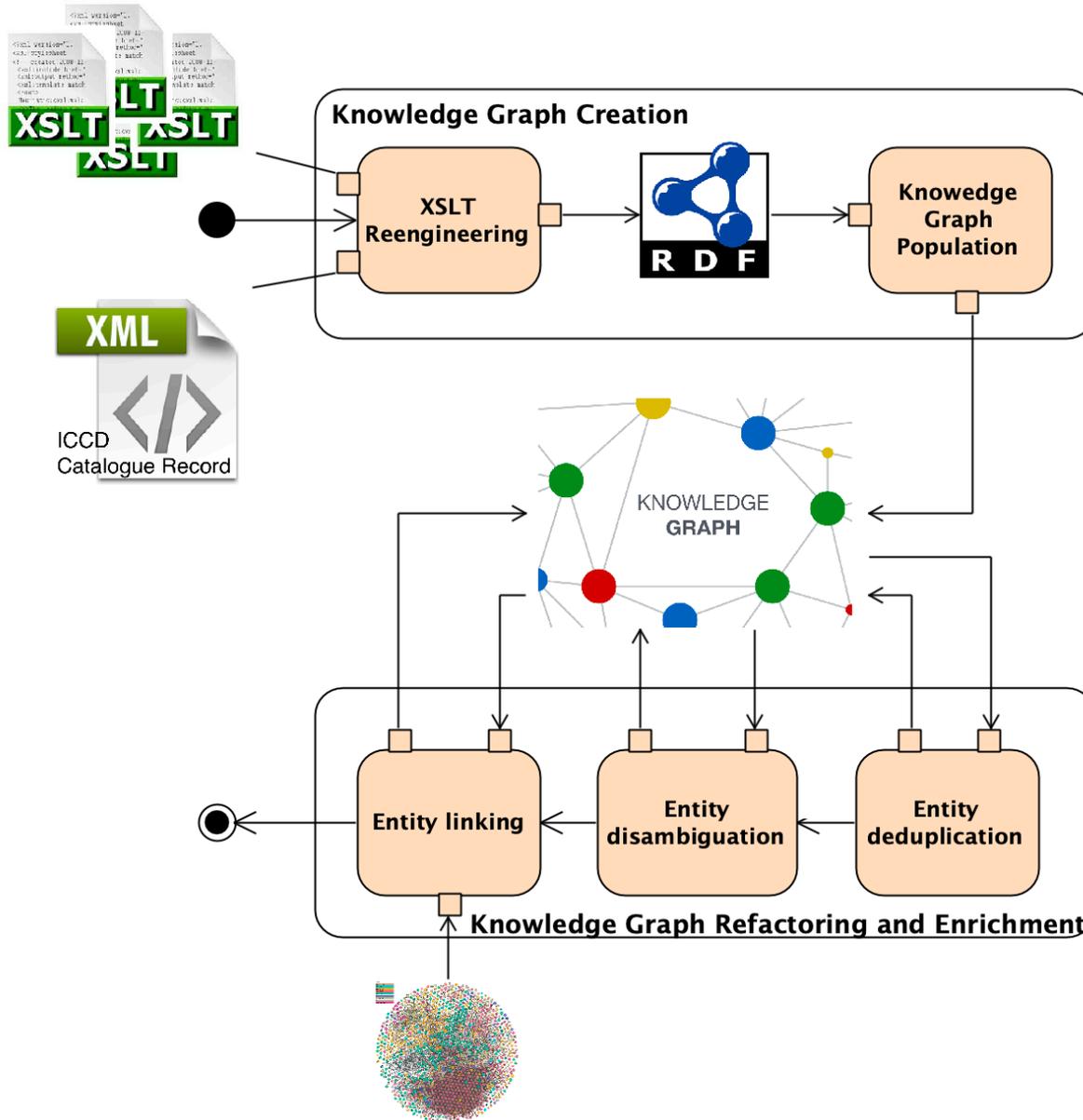
arco:MovableCulturalProperty

↳ Bene mobile

Convertire i dati con RDFizer

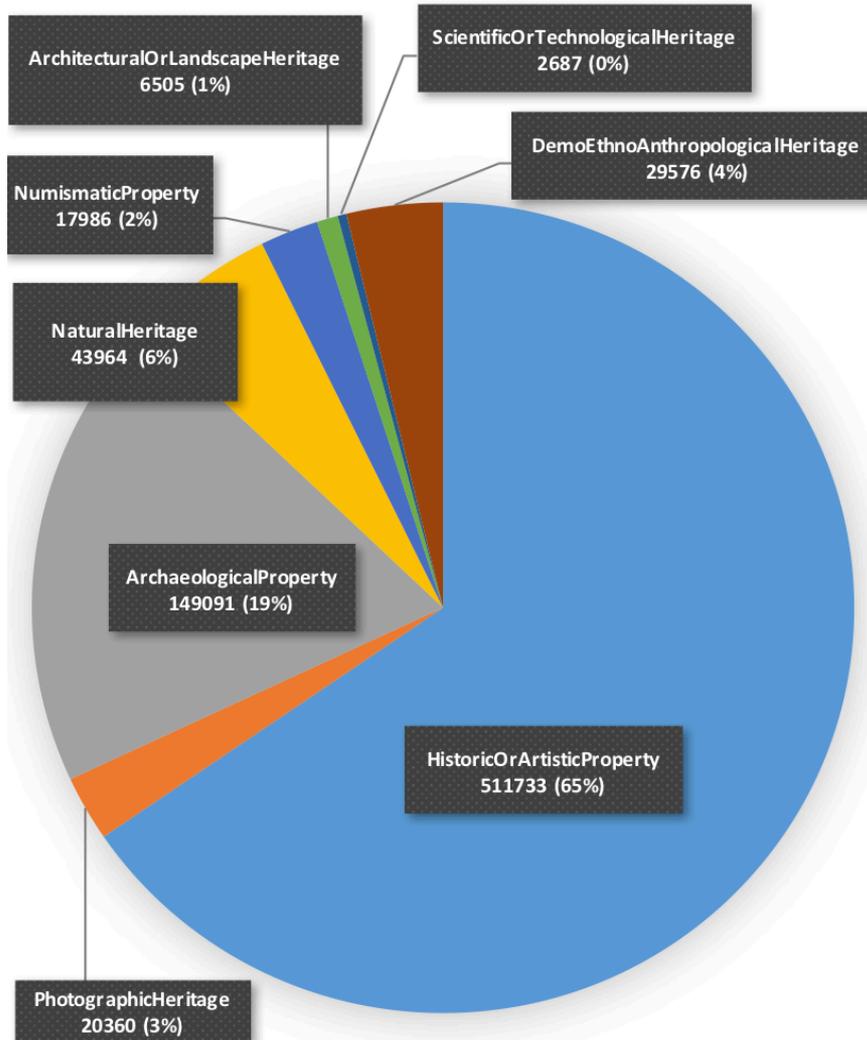
- Software di conversione dei dati XML rappresentati secondo gli standard di catalogazione ICCD in RDF modellati secondo le ontologie di ArCo
- Implementato in Java 8 e supporto Maven
- Open Source sul repo GitHub del progetto



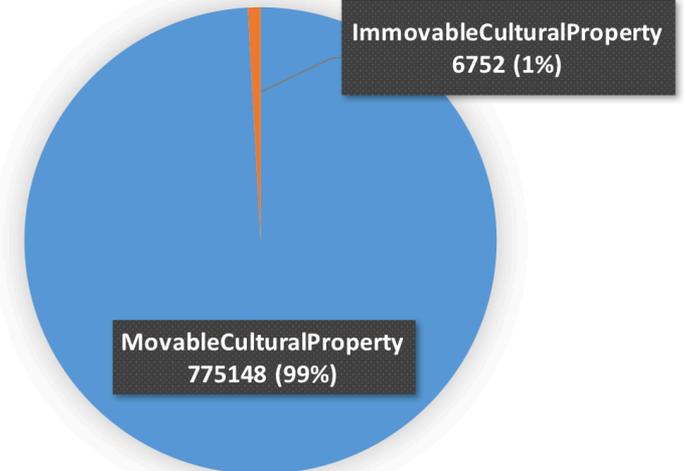


Un po' di numeri

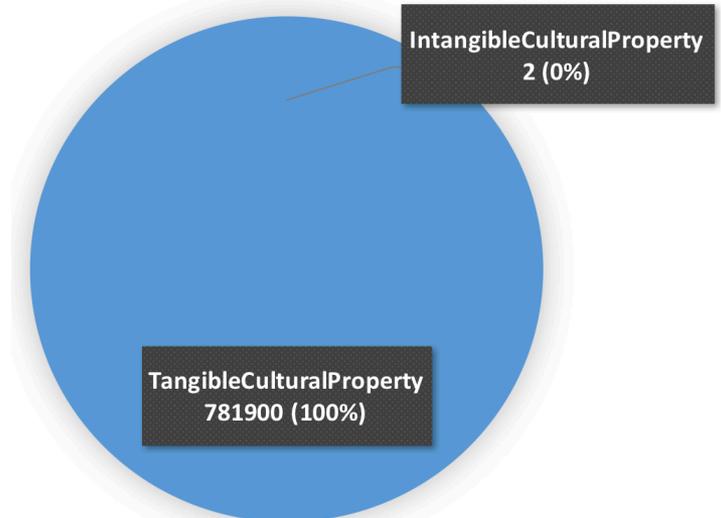
Beni culturali per tipo



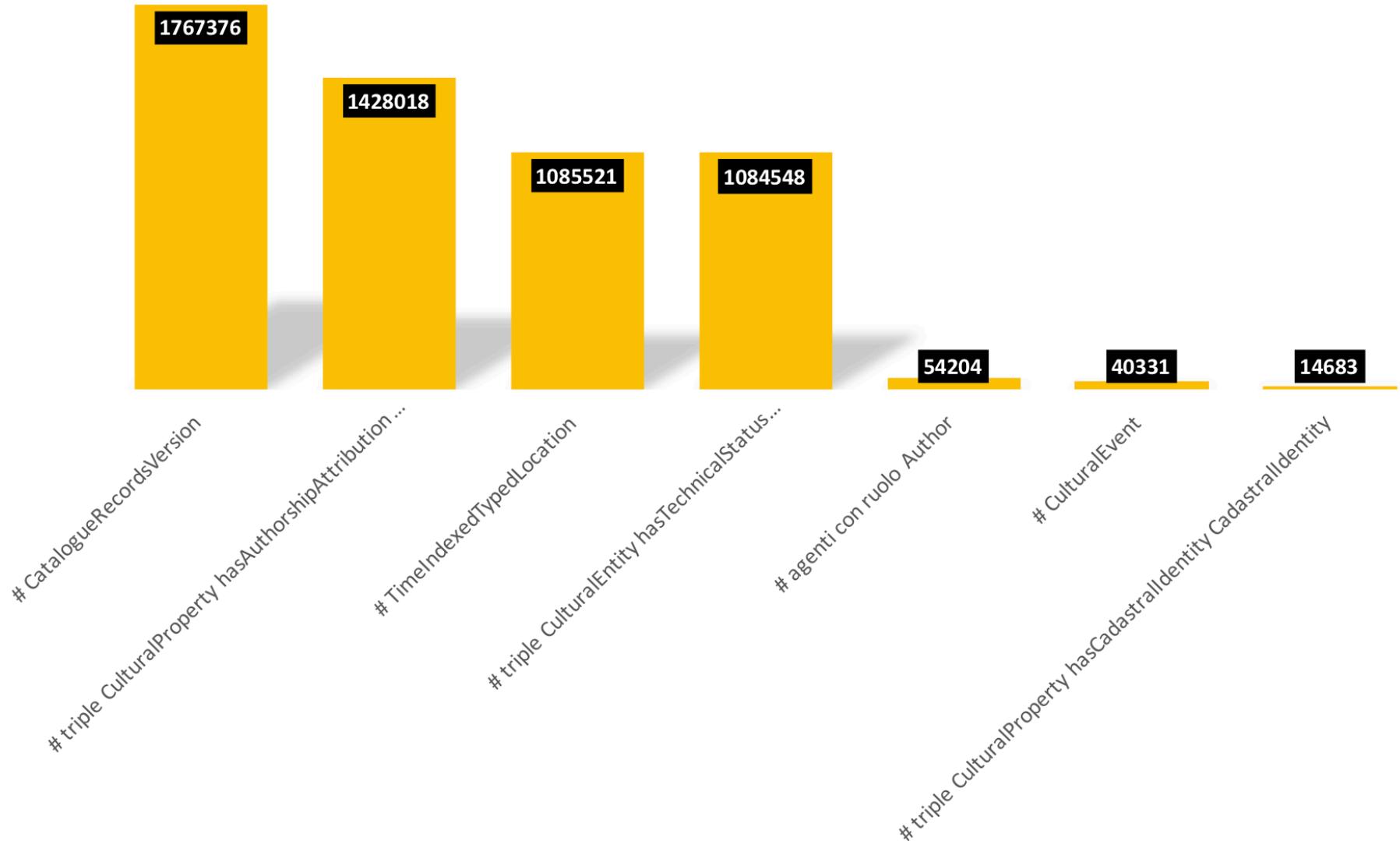
ImmovableCulturalProperty Vs. MovableCulturalProperty



Tangible Vs. Intangible



Esempi di uso di pattern nei dati



Scelte di data design

Always be FAIR!

- **Findable:**
 - URI permanenti (w3id) per identificare entità sia nelle ontologie che nei dati
 - DOI (10.5281/zenodo.2630447) per identificare/citare il Knowledge Graph di ArCo
- **Accessible:** ci basiamo su protocolli standard aperti (HTTP(S)) per l'accessibilità e l'interrogazione (SPARQL)
- **Interoperable:** usiamo protocolli standard aperti per la modellazione della conoscenza (RDF e OWL)
- **Reusable:** il Knowledge Graph è rilasciato con licenza CC BY-SA 4.0

URI permanenti

- Le URI permanenti sono costruite a partire dal base URI <https://w3id.org/arco/>
- Ontologie
 - Base namespace:
<https://w3id.org/arco/ontology/{ID Ontologia}>
 - Esempio: <https://w3id.org/arco/ontology/core>
- Dati
 - Base namespace:
<https://w3id.org/arco/resource/{Tipo}/{ID entità}>
 - Esempio:
<https://w3id.org/arco/resource/Agent/dcd4ca7b54dd3d7dac083dd4c54a9eef>

Deduplicazione e disambiguazione

Politica per generazione ID entità

1. Identifichiamo un set di elementi nel sorgente XML che costituiscono una possibile chiave
2. Rimuoviamo i caratteri non ammessi in una URI e convertiamo il valore in lower case
3. Riordiniamo in ordine alfabetico
4. Calcoliamo l'MD5 checksum sulla stringa risultante

<AUTN>Frischia Albert</AUTN>



friscia-albert



ex. albert-friscia



dcd4ca7b54dd3d7dac083dd4c54a9eef.

<https://w3id.org/arco/resource/Agent/dcd4ca7b54dd3d7dac083dd4c54a9eef>

Deduplicazione

- Collassare individui distinti generati per la stessa entità
 - Ex: “Andrea d'Agnolo” e “Andrea d'Agnolo detto del Sarto”
- Deduplicazione ricondotta ad un problema di allineamento interno
 - ArCo utilizzato sia come dataset sorgente che target
- La deduplicazione ottenuta con LIMES
 - Metrica per il linking: Jaccard distance sui valori rdfs:label associati alle entità
 - Threshold estremamente selettivo: 0.8 sul range [0-1]

Disambiguazione

- Disaccoppiare individui erroneamente collassati su una stessa entità
- Studio di motif che forniscono l'impronta digitale di un'entità
 - Ex. Periodo di attività di un autore e tipo di beni ad egli associati
- Attività in corso
- Ci aspettiamo un numero ristretto di casi rispetto alle dimensioni del Knowledge Graph

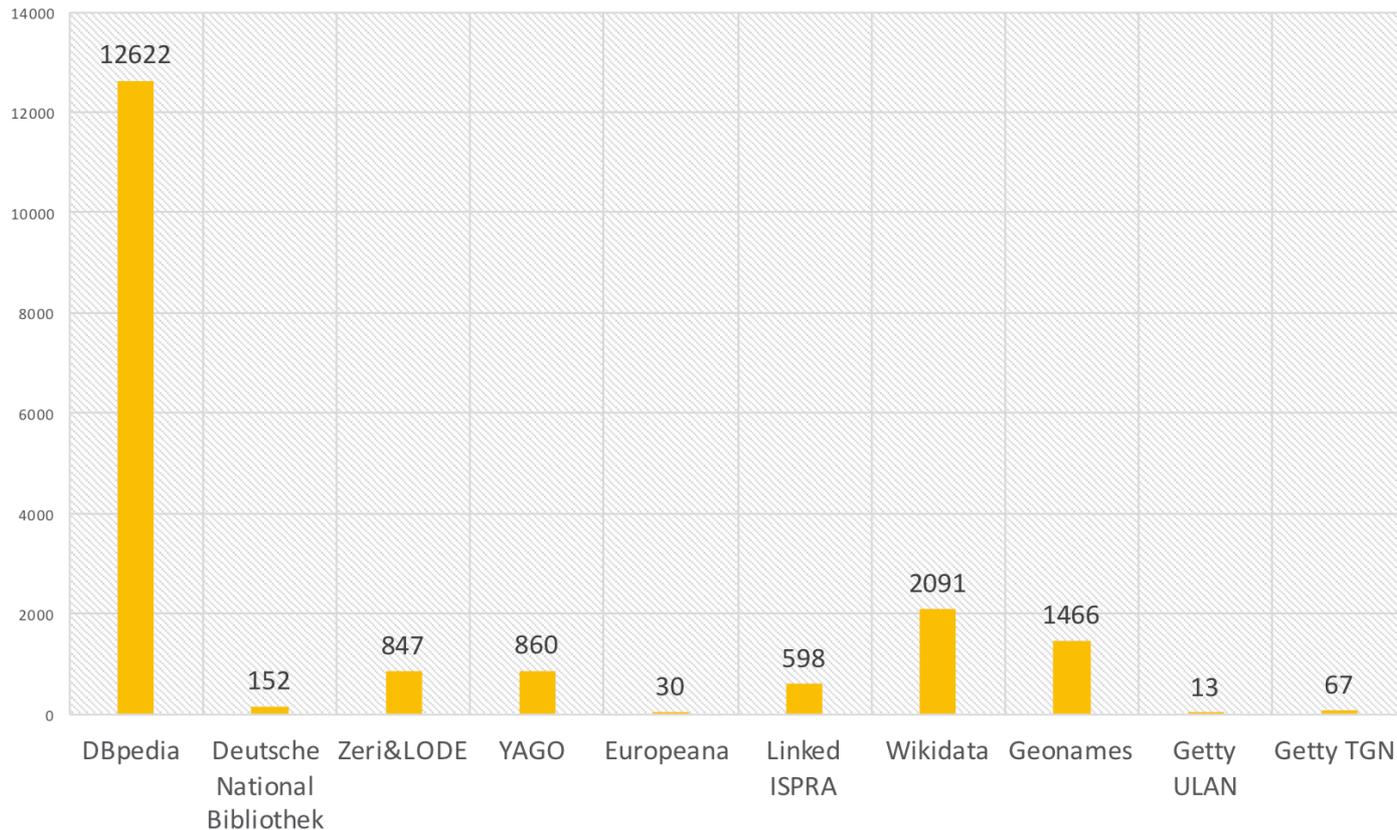
Entity linking

LIMES - Linking

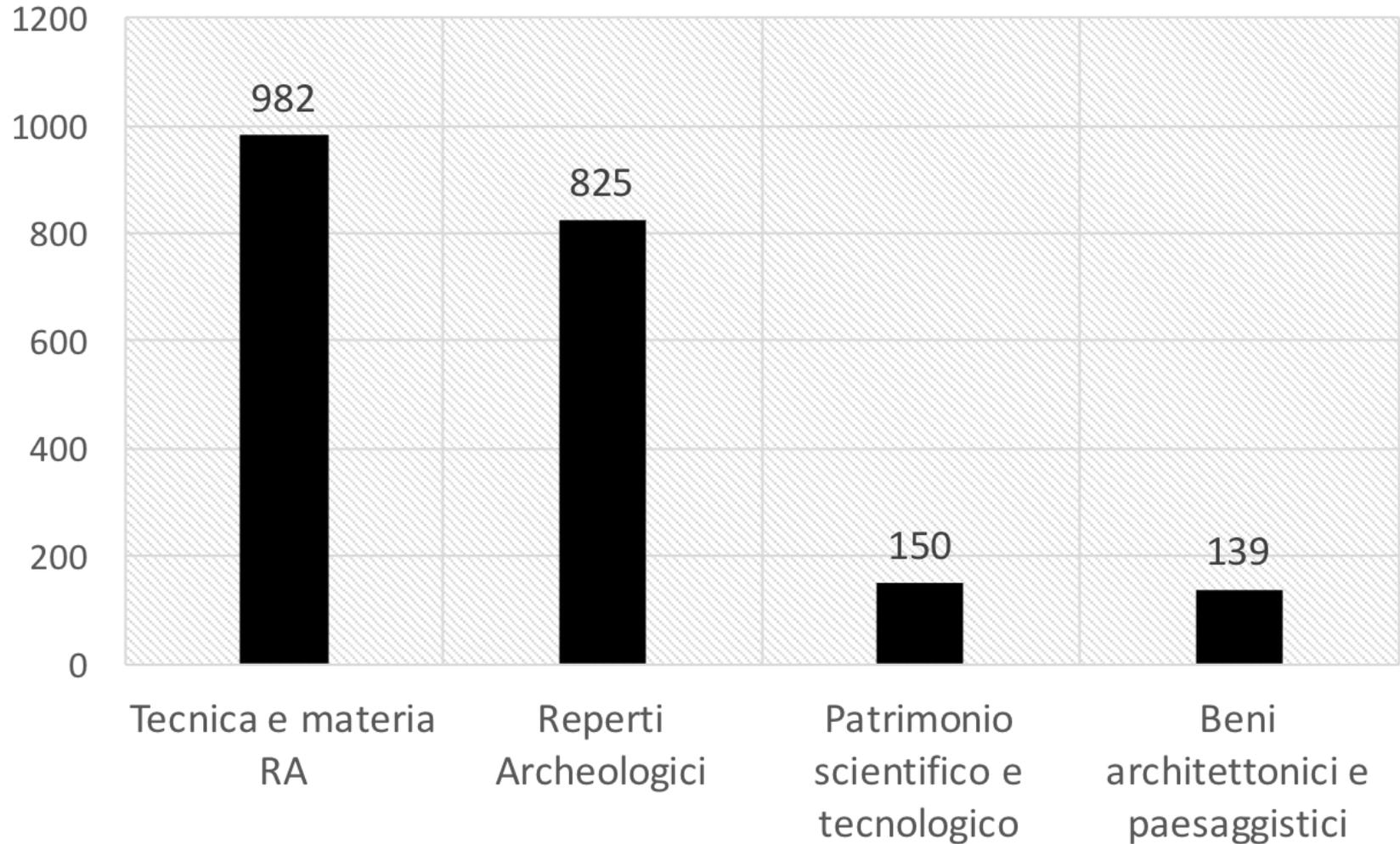
- L'entity linking è ottenuto utilizzando LIMES
- Metrica per il linking: Jaccard distance sui valori rdfs:label associati alle entità
- Threshold estremamente selettivo: 0.9 sul range [0-1]
 - Al di sotto del threshold i link candidati sono scartati
 - 0.9 approssima risultati affidabili al 100% (test manuale su un campione del 10% dei link ottenuti)
- Link rappresentati con assiomi owl:sameAs

Tipi di entità e LOD collegati

- Collegati autori (8.884 link) e luoghi (9.862 link)
- Dataset collegati



Vocabolari controllati dati.beniculturali.it

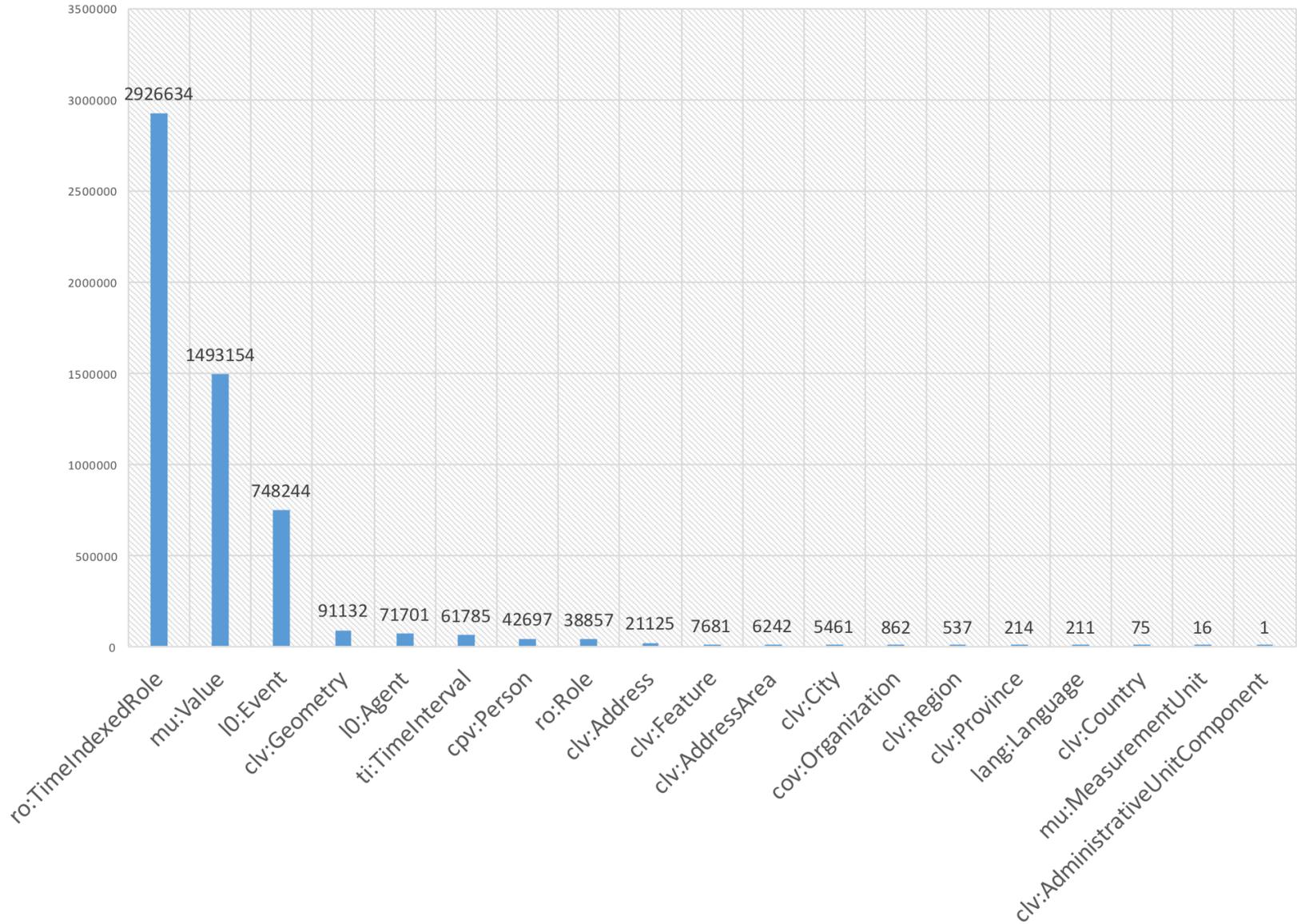


ArCo e OntoPiA

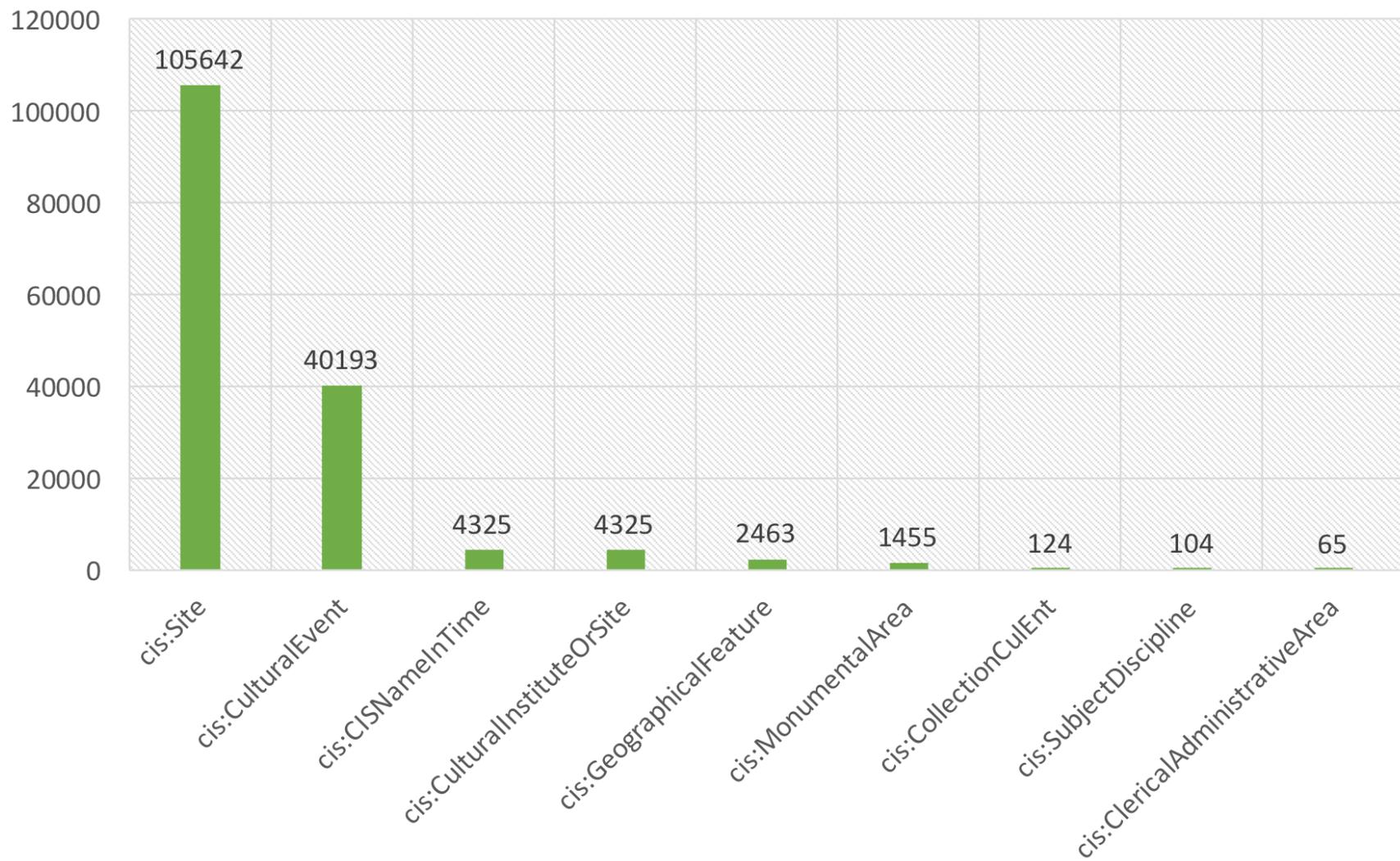
ArCo è parte di OntoPiA

- OntoPiA-L0 usata come top-level ontology di ArCo
- ArCo riusa direttamente
 - Level-0 (I0),
 - Role ontology (RO)
 - Measurement Unit Ontology (MU)
 - Core Location Vocabulary (CLV)
 - Time ontology (TI)
 - Core Person Vocabulary (CPV)
 - Core Organization Vocabulary (COV)
 - Language Ontology
 - Cultural-ON

Riuso di OntoPiA in numeri



Riuso di Cultural-ON in numeri



Docker container di ArCo

ICCD-MiBACT / ArCo

Unwatch 12 Star 16 Fork 1

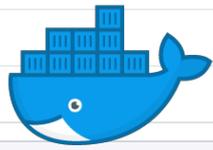
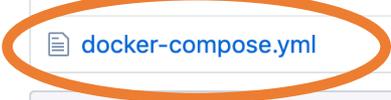
Code Issues 9 Pull requests 0 Projects 0 Wiki Insights

Branch: master ArCo / ArCo-release /

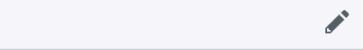
Create new file Upload files Find file History

ValentinaPasqual Update index-it.html Latest commit 60057a4 7 hours ago

..		
httpd	Update index-it.html	7 hours ago
lode	Change name of the release.	a year ago
ontologie	refactoring	a day ago
rdfizer	rdfizer version 1.0	21 hours ago
test	Update testcase-26b.owl	4 days ago
virtuoso	Release v0.5	a month ago
README.md	minor	a year ago
docker-compose.yml	Release 0.4	4 months ago



docker



README.md

ArCo





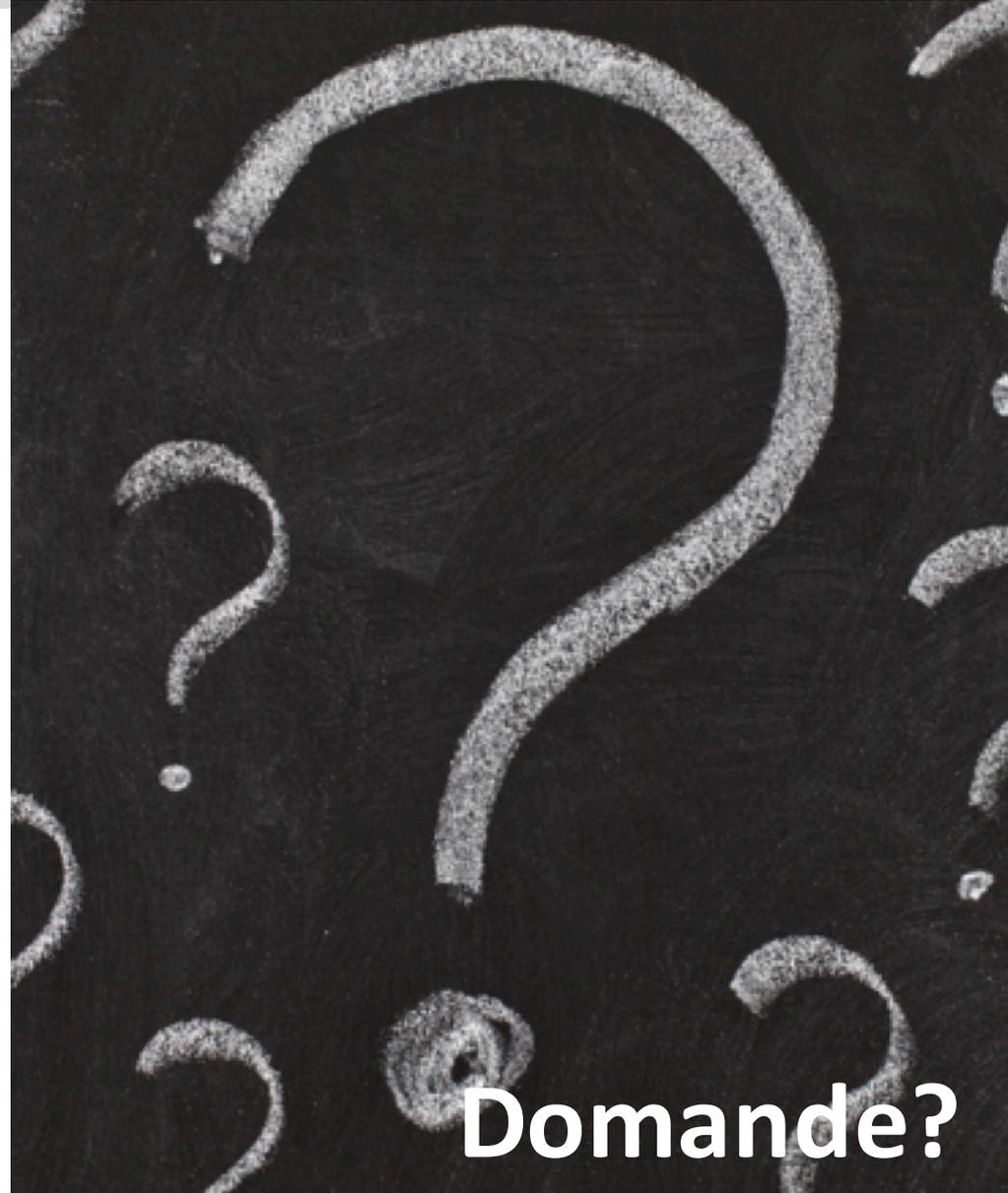
Valentina Carriero
STLab, ISTC-CNR

valentina.carriero@istc.cnr.it



Andrea Giovanni Nuzzolese
STLab, ISTC-CNR

andrea.nuzzolese@cnr.it



Domande?