



Dal data model DAAS ai Linked Open Data

*Mappatura del data model DAAS (ICCU/MiC) verso gli standard del
web semantico*

1. Il punto di partenza

Il materiale di partenza è il **dump DAAS**: una grande esportazione di dati in formato **JSONL** (un record JSON per riga), per un totale di circa **22,2 milioni di record di testa** e **~200 GB**. I dati sono organizzati in tre domini, identificati dai rispettivi namespace interni, con volumi molto diversi tra loro:

Dominio	Namespace	Record di testa	Quota	Cosa contiene
Bibliografico	:bibl	21.900.439	98,5 %	opere, edizioni, periodici SBN e tutto il loro contorno (identificativi, responsabilità, biblioteche, datazioni)
Agenti / protagonisti	:abap :arch :bibl	203.111	0,9 %	persone, enti, famiglie come authority a sé stanti
Archivistico	:arch	120.082	0,5 %	soggetti produttori d'archivio (enti, persone, famiglie)

Una difficoltà iniziale: al ricevimento del dump **non esisteva una documentazione formale del modello dati**. La struttura — quali entità esistono, come sono collegate, quali attributi portano — è stata quindi **dedotta empiricamente** leggendo il contenuto reale dei record.

L'obiettivo del lavoro è **pubblicare questi dati come Linked Open Data**, cioè in RDF, usando URI di **ontologie pubbliche, risolubili e standard** in modo che i dati siano interoperabili con l'ecosistema dei beni culturali (a partire da quello italiano) e con il web semantico in generale.

2. Il percorso di analisi (7 step)

Il lavoro ha seguito sette tappe. Ogni decisione è stata **verificata contro le ontologie reali** — controllando tipo OWL, dominio e codominio di ogni termine.

Step 1 — Estrazione del data model. Leggendo i ~22 milioni di record (in parallelo, per ragioni di volume) abbiamo ricostruito lo schema implicito del dump: le classi effettivamente istanziate, le relazioni tra classi e gli attributi testuali, ciascuno con la sua frequenza d'uso.

Step 2 — Copertura con il filone preferito. Per ogni classe e proprietà ci siamo chiesti: è già esprimibile col filone preferito (ArCo + OntoPiA)? La risposta non era sempre diretta, perché il modello concettuale di riferimento (I.PaC) organizza i termini in una **gerarchia**: spesso un termine specifico eredita il mapping da una sua superclasse. Risalendo questa gerarchia, e validando ogni target proposto contro 19 ontologie ufficiali, abbiamo coperto la maggior parte del modello — e scoperto anche

alcune **incoerenze nascoste** (proprietà usate in modo diverso da come le definisce l'ontologia di destinazione).

Step 3 — Fallback. Per ciò che il filone preferito non copriva, abbiamo cercato l'equivalente negli standard di fallback: **BIBFRAME** per il bibliografico, **Arco4Archives** per l'archivistico. È emerso che il dominio archivistico era **già interamente coperto** dal filone preferito (OntoPiA), e che Arco4Archives non è stato necessario.

Step 4 — Punti aperti (IN PROGRESS). Abbiamo raccolto i termini per cui il mapping **non era univoco**: voci “orfane” (senza target negli Excel di mapping fornito) e “anomalie” (termini con un target ma incoerente). Per ognuna abbiamo proposto un target **già validato** e formulato la domanda precisa da portare al committente.

Step 5 — Estensioni “Protagonisti” Per Portale (IN PROGRESS). In parallelo abbiamo valutato una serie di requisiti di arricchimento per la sezione Protagonisti del Portale.

3. La strategia: una gerarchia di preferenza

La regola guida, concordata con il committente, è una **gerarchia di preferenza** sugli standard di destinazione. Per ogni termine si sceglie il primo che lo esprime adeguatamente:

1. `:cross` → ArCo + OntoPiA + Cultural-ON (**preferito**) — **l'ecosistema ufficiale dei beni culturali italiani, ancorato al web semantico della PA (OntoPiA)**;
2. `:bibl` → BIBFRAME (**fallback**) — **quando il concetto è propriamente bibliografico e non ha un equivalente nel filone preferito (es. edizione, descrizione fisica, serialità)**;
3. `:arch` → ArCo4Archives (**fallback**) — **per l'archivistico, non risultato necessario perché già coperto da OntoPiA.**

A questo si aggiunge una **regola sugli URI delle risorse**: gli identificativi degli agenti che già esistono nel sistema ICCD (riconoscibili dal segmento /SCN/) vengono ricondotti agli URI ArCo ufficiali (<https://w3id.org/arco/resource/...>); tutto il resto riceve un URI sotto un namespace dedicato (<https://w3id.org/italia/cultural-heritage/resource/...>).

Un principio importante: quando un'entità ha **più nature** — per esempio è insieme un bene culturale e una manifestazione bibliografica — non si sceglie l'una o l'altra, ma si adottano **entrambe** (vedi il pattern “doppia tipizzazione”, §8).

4. Il data model risultante

Il modello finale **non introduce termini propri**: riusa termini di ontologie pubbliche, dichiarando per ciascuno l'ontologia di origine. È un profilo applicativo — l'insieme dei termini standard che il dump DAAS effettivamente popola.

Ecosistema	Ontologie	A cosa serve	Esempi di termini
ArCo	core, arco-lite, context-description	beni culturali italiani (identificativi, responsabilità, datazioni)	core:Identifier, a-cd:Responsibility, a-cd:agentDate
OntoPiA	l0, CPV, COV, CLV, TI, SM, RO	entità di base della PA: persone, enti, luoghi, tempo	l0:Agent, CPV:Person, COV:Organization, CLV:Address, TI:TemporalEntity, SM:URL
Cultural-ON	cis	l'entità culturale e i suoi luoghi	cis:CulturalEntity, cis:Site
BIBFRAME	bf	il bibliografico (manifestazioni, edizione, serialità)	bf:Instance, bf:editionStatement, bf:continues
FRBRoo	frbroo	personaggi della finzione	frbroo:Character
IFLA / RDA	ISBD, RDA-U	aree ISBD, punti d'accesso autorizzati	isbdu:P1163, rdau:P61170
altri vocabolari	SKOS, FOAF, DCTerms	etichette, genere, note generiche	skos:prefLabel, foaf:gender

Copertura raggiunta:

- **23 classi su 23** mappate (le 23 classi distinte istanziate nel dump, che compaiono nei tre domini in forma namespace-qualificata — es. :bibl/Person, :abap/Person, :arch/Person sono la stessa classe CPV:Person);
- **28 object property su 28** mappate;
- **34 datatype property: 29 mappate direttamente + 3 reificate** su nodo tipizzato (relationType, nameType, issuedBy) + **2 in sospeso** (type, temporalEntityType), che sono scelte di modellazione da ratificare, non target mancanti (§10).

Il crosswalk puntuale termine-per-termine è al §7. Il modello machine-readable è fornito anche come profilo RDF (data-model.ttl), dove ogni termine dichiara l'ontologia di origine (rdfs:isDefinedBy) e la provenienza nel dump DAAS (skos:note); non conia alcun termine proprio e non importa le ontologie d'origine.

5. Anatomia di una risorsa bibliografica (esempio con dati fittizi)

Un'opera bibliografica è tutto ciò che le ruota intorno. Gli identificativi sono inventati per leggibilità:

```
ex:opera-1 a cis:CulturalEntity, bf:Instance ;           # doppia natura: culturale + bibliogr.
  arco-lite:primaryName "Saper vedere l'arte" ;
  bf:editionStatement "2. ed." ;                         # ISBD area 2 – edizione
  bf:provisionActivityStatement "Milano : Esempi , 1960" ; # ISBD area 4
  isbdu:P1163 "254 p. : ill. ; 23 cm" ;                 # ISBD area 5 – descr. fisica
  core:hasIdentifier ex:id-1 ;                           # identificativo primario
  arco-lite:alternativeIdentifier "ALT-99" ;              # identificativo alternativo (literal)
  a-cd:hasResponsibility ex:resp-1 ;                     # chi ne è responsabile
  arco-lite:hasOwner ex:biblioteca-1 ;                   # chi la possiede
  arco-lite:hasRealizationDate ex:tempo-1 .              # quando è stata pubblicata

ex:id-1 a core:Identifier ; core:identifier "SBN-12345" ;
  core:hasType ex:tipo-isbn .                             # il TIPO (schema)
ex:tipo-isbn a core:IdentifierType ; rdfs:label "ISBN" .

ex:resp-1 a a-cd:Responsibility ;
  a-cd:hasAgentWithResponsibility ex:persona-1 ;         # l'agente: Mario Rossi
  core:hasRole ex:ruolo-autore ;                         # il ruolo: Autore
  a-cd:hasResponsibilityType a-cd:PrimaryResponsibility . # il grado: primaria

ex:persona-1 a CPV:Person ; arco-lite:primaryName "Rossi, Mario" .
ex:tempo-1 a TI:TemporalEntity ; TI:date "1960" .
```

Come si legge: un'**opera** ha un'**identità doppia** (culturale e bibliografica), porta i propri dati catalografici ISBD, ed è collegata a chi ne è **responsabile** (Mario Rossi, in qualità di autore, con responsabilità primaria), alla **biblioteca** che la possiede e al **tempo** di pubblicazione. Ogni “cosa” (l'opera, la persona, la biblioteca, l'identificativo) è un nodo con un proprio URI, riutilizzabile e collegabile ad altri dati.

6. I tre domini in pratica

Il bibliografico (§5) è il caso più ricco; gli altri due domini condividono lo stesso vocabolario di base (OntoPiA per persone/enti/famiglie, ArCo per gli identificativi).

6.1 Agenti / protagonisti (:abap)

Authority di persone, enti e famiglie a sé stanti. Le classi istanziate sono Agent (→ l0:Agent), Person (→ CPV:Person), Organization (→ COV:Organization), Family (→ CPV:Family), con annidati Identifier, Feature (luogo di nascita), Site.

```
ex:persona-2 a CPV:Person ;                               # una specializzazione di l0:Agent
  arco-lite:primaryName "Cagnetto, Giovanni" ;
  rdau:P61170 "Cagnetto, Giovanni <1875-1939>" ; # punto d'accesso autoriz.
  CPV:givenName "Giovanni" ;
  CPV:dateOfDeath "1939-12-24" ;
  CPV:hasBirthPlace ex:luogo-1 ;                         # → CLV:Feature
  core:hasIdentifier ex:id-isni .
ex:luogo-1 a CLV:Feature ; l0:name "Napoli" .
ex:id-isni a core:Identifier ; core:identifier "0000000121012345" ;
  core:hasType ex:tipo-isni .
ex:tipo-isni a core:IdentifierType ; rdfs:label "ISNI" .
```

Una **verifica indipendente** sul dump AGENT standalone (scansione **completa**: 6.702.036 record) ha confermato che **il 100 % delle classi, object property e datatype property del dominio agenti ha un target nel modello** — nessuna voce scoperta.

6.2 Archivistico (:arch)

Soggetti produttori d'archivio: Organization (→ COV:Organization), Person (→ CPV:Person), Family (→ CPV:Family), con Identifier e TemporalEntity (estremi cronologici di attività). Strutturalmente è anch'esso un insieme di agenti, e per questo — come emerso allo step 3 — è **già interamente coperto da OntoPiA**, senza bisogno del fallback RiC-O.

```
ex:ente-arch a COV:Organization ;
  arco-lite:primaryName "Comune di Esempio" ;
  rdau:P61170 "Comune di Esempio" ;
  arco-lite:hasRealizationDate ex:tempo-2 ; # estremi cronologici
  core:hasIdentifier ex:id-arch .
ex:tempo-2 a TI:TemporalEntity ; TI:startTime "1860" ; TI:endTime "1927" .
ex:id-arch a core:Identifier ; core:identifier "ARCH-7788" .
```

7. Crosswalk completo DAAS → LOD

Tabelle complete classe/proprietà → target standard, con il **volume** (numero di occorrenze osservate nel dump). Sono la versione esaustiva del modello; rispecchiano 1:1 ciò che la pipeline di conversione emette.

7.1 Classi (23)

Classe dump	Volume	Target
Identifier	327.633.413	core:Identifier
Site	110.591.480	cis:Site
Library	110.587.605	COV:Organization
Address	109.885.882	CLV:Address
Responsibility	32.577.045	a-cd:Responsibility
Type	30.228.366	core:Type
Role	23.810.725	R0:Role
Name	23.235.542	a-cd:DesignationInTime
Person	19.681.949	CPV:Person
Organization	18.307.998	COV:Organization
WorkInstance	18.085.032	cis:CulturalEntity (+ bf:Instance)
TemporalEntity	16.242.881	TI:TemporalEntity
BibliographicLevel	15.025.858	bf:Issuance
Series	4.123.534	cis:CulturalEntity
Subject	3.535.424	a-cd:Subject
Classification	2.260.300	bf:Classification

Classe dump	Volume	Target
Work	1.816.330	cis:CulturalEntity
Meeting	290.221	bf:Meeting
Agent	193.735	l0:Agent
Character	100.956	frbroo:Character
Feature	24.913	CLV:Feature
Family	7.069	CPV:Family
Manuscript	35	cis:CulturalEntity

7.2 Object property (28)

Property dump	Volume	Target	Note
hasAlternativeIdentifier	196.944.182	arco-lite:alternativeIdentifier	appiattito a literal (modello duale, §8.1)
hasPrimaryIdentifier	130.689.209	core:hasIdentifier	
hasSite	110.591.480	CLV:hasSpatialCoverage	
hasOwner	110.568.161	arco-lite:hasOwner	
hasAddress	109.885.882	cis:siteAddress	
hasCulturalEntity	32.577.045	a-cd:hasResponsibility	
hasResponsibility	31.403.598	a-cd:hasAgentWithResponsibility	
hasType	30.228.366	core:hasType	
hasRole	23.810.725	core:hasRole	
hasPreferredName	21.833.692	a-cd:hasDesignationInTime	
hasPublicationOrRealizationTime	15.215.848	arco-lite:hasRealizationDate	
hasBibliographicLevel	15.025.858	bf:issuance	
isMemberOf	4.123.534	arco-lite:isCollectionMemberOf	
hasSubject	3.535.424	a-cd:hasSubject	
hasPart	2.778.103	core:hasPart	
hasClassification	2.260.300	bf:classification	
hasWorkInstance	1.816.330	bf:hasInstance	
hasAlternativeName	1.401.850	a-cd:hasDesignationInTime	
hasTemporalCoverage	1.027.033	TI:hasTemporalCoverage	
contains	154.957	bf:hasPart	relazione seriale
hasSupplement	64.012	bf:supplement	relazione seriale
continues	47.065	bf:continues	relazione seriale
hasBirthPlace	24.913	CPV:hasBirthPlace	

Property dump	Volume	Target	Note
hasOtherEdition	9.017	bf:otherEdition	relazione seriale
continuesInPart	3.721	bf:continuesInPart	relazione seriale
mergesWith	1.294	bf:mergedToForm	relazione seriale (approssimata, §10)
absorbs	1.040	bf:absorbed	relazione seriale
hasIdentifier	22	core:hasIdentifier	

7.3 Datatype property (34)

Property dump	Volume	Target	Note
originalId	327.538.024	core:identifier	
authorityLabel	148.745.437	rdau:P61170	
fullAddress	109.885.882	CLV:fullAddress	
prefLabel	71.325.249	skos:prefLabel	
preferredName	55.238.312	arco-lite:primaryName	
code	47.514.524	skos:notation	
type	26.923.603	core:hasType → individuo	
name	23.260.373	l0:name	
relationType	18.037.264	a-cd:hasResponsibilityType → individuo	
date	14.574.184	TI:date	
isbdArea5PhysicalDescription	14.522.444	isbdu:P1163	
isbdArea4PublicationDistribution	14.438.240	bf:provisionActivityStatement	
temporalEntityType	11.860.510	core:hasType → individuo	
description	8.307.567	l0:description	
note	5.139.204	core:note	
issuedBy	5.105.542	core:issuedBy → nodo	
givenName	4.346.623	CPV:givenName	
isbdArea2Edition	1.395.958	bf:editionStatement	
time	1.120.628	TI:time	
endDate	547.951	TI:endTime	
startDate	547.951	TI:startTime	
dateOfDeath	488.319	CPV:dateOfDeath	
isbdArea3SpecificTypeOfResource	213.932	isbdu:P1161	
nameType	159.054	core:hasType → individuo	reificato (riuso ArCo)
identifier	94.073	core:identifier	

Property dump	Volume	Target	Note
agentDate	77.792	a-cd:agentDate	
abstract	75.537	rico:generalDescription	
localIdentifier	71.332	arco-lite:localIdentifier	
sexInformation	44.197	foaf:gender	
label	24.913	l0:name	
contentsNote	24.885	rico:generalDescription	
url	17.729	SM:URL	
historicalBiographicalInformation	1.006	a-cd:historicalBiographicalInformation	
toponym	1	l0:name	

8. I pattern di modellazione

Lavorando sui dati reali sono emersi alcuni **modi ricorrenti** con cui i termini si compongono. Riconoscerli rende il modello prevedibile e documentabile. Ogni pattern ha una **regola formale** verificabile (SHACL shape, §9). Esempi a dati fittizi.

8.1 Identificativo (modello duale)

Un'entità può avere un identificativo primario e diversi alternativi. Il primario è modellato come **nodo ricco** (porta ente emittente, tipo...), gli alternativi come semplici **valori testuali**, per non appesantire il grafo.

```
ex:opera-1 core:hasIdentifier ex:id-1 ;           # primario → nodo
          arco-lite:alternativeIdentifier "ALT-99" . # alternativo → literal
ex:id-1 a core:Identifier ;
      core:identifier "SBN-12345" ;               # obbligatorio
      core:issuedBy ex:org-emittente ;           # opzionale
      core:hasType ex:tipo-isbn .                 # opzionale (lo schema, §10)
```

Shape: core:identifier minCount 1; issuedBy → core:Organization; hasType → core:Type.

8.2 Responsibility (chi ha fatto cosa)

Collega un'entità culturale a un agente, specificando **ruolo** (autore, editore...) e **grado** di responsabilità (primaria, secondaria...). È un nodo intermedio che reifica il legame.

```
ex:opera-1 a-cd:hasResponsibility ex:resp-1 .
ex:resp-1 a a-cd:Responsibility ;
      a-cd:hasAgentWithResponsibility ex:persona-1 ; # → Person / Organization / Meeting
      core:hasRole ex:ruolo-editore ;                 # → RO:Role ("Editore")
      a-cd:hasResponsibilityType a-cd:SecondaryResponsibility .
```

Shape: hasResponsibilityType vincolato (sh:in) ai 4 individui ArCo Primary/Secondary/Coordinated/AlternativeOrOutdated.

8.3 Tipo controllato (literal → individuo ⊆ core:Type)

Spesso il dump contiene una **stringa** dove il modello vuole un **valore di vocabolario controllato**. La stringa viene ricondotta all'individuo standard corrispondente.

```
ex:resp-1 a-cd:hasResponsibilityType a-cd:PrimaryResponsibility . # dalla stringa "Primaria"
ex:nome-1 core:hasType a-cd:ParallelTitle . # dalla stringa "parallelo"
```

8.4 Doppia tipizzazione

Quando una risorsa ha più nature, le si assegnano **più tipi** contemporaneamente. La WorkInstance è insieme cis:CulturalEntity (bene culturale) e bf:Instance (manifestazione bibliografica): la seconda natura è ciò che rende domain-coerenti le aree ISBD (che hanno domain bf:Instance). Lecito: non c'è disjointWith tra le due classi.

```
ex:opera-1 a cis:CulturalEntity, bf:Instance .
```

8.5 Nome / Designazione

Nomi e titoli (preferiti, alternativi, paralleli) sono modellati come nodi “designazione”, con il testo e l'eventuale tipo.

```
ex:opera-1 a-cd:hasDesignationInTime ex:titolo-1 .
ex:titolo-1 a a-cd:DesignationInTime ; l0:name "Saper vedere l'arte" . # l0:name esattamente 1
```

8.6 Luogo (catena sede → indirizzo) e Feature geografica

Un ente possiede una sede, che ha un indirizzo; il luogo di nascita di una persona è una Feature geografica.

```
ex:biblioteca-1 CLV:hasSpatialCoverage ex:sede-1 .
ex:sede-1 a cis:Site ; cis:siteAddress ex:indirizzo-1 .
ex:indirizzo-1 a CLV:Address ; CLV:fullAddress "Via Esempio 1, Roma" . # obbligatorio
ex:luogo-1 a CLV:Feature ; l0:name "Napoli" . # obbligatorio
```

8.7 Tempo

Date di pubblicazione/realizzazione e intervalli.

```
ex:opera-1 arco-lite:hasRealizationDate ex:tempo-1 .
ex:tempo-1 a TI:TemporalEntity ; TI:date "1960" . # oppure TI:startTime / TI:endTime
```

8.8 Relazioni seriali

Tra periodici, con il **tipo esplicito** della relazione (non un generico bf:relation).

```
ex:rivista-2024 bf:continues ex:rivista-1990 ;    # continua una testata precedente
                bf:absorbed   ex:bollettino .      # ne ha assorbita un'altra
```

9. Garanzia di qualità: la validazione SHACL

Per ciascun pattern abbiamo scritto delle **regole formali verificabili** (SHACL shapes), derivate empiricamente dai dati reali (cardinalità e tipi osservati su migliaia di istanze). Esempi di regola:

- ogni identificativo deve avere un valore (`core:identifier minCount 1`);
- una responsabilità ha un grado tra i quattro ammessi (primaria / secondaria / coordinata / alternativa o sorpassata);
- chiunque porti un'area ISBD deve essere anche una `bf:Instance` (oltre che `cis:CulturalEntity`) — cioè la doppia tipizzazione è obbligatoria, non opzionale.

Queste regole funzionano da **contratto**: ogni file RDF prodotto dalla conversione viene validato automaticamente e, se non conforme, segnalato. Durante il lavoro la validazione ha già intercettato output non aggiornati, dimostrandone l'utilità come rete di sicurezza. È la base del controllo qualità sull'intero output finale.