JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

JLIS.it

Entity modeling: traces of an evolving path

Tiziana Possemato[a]

a) @Cult, https://orcid.org/0000-0002-7184-4070

Contact: Tiziana Possemato, tiziana.possemato@atcult.it
Received: 26 May 2022; **Accepted:** 3 June 2022; **First Published:** 15 September 2022

ABSTRACT
In this work we will deal with the subject of entities understood as real-world objects and how this concept is used in the context of entity modeling, that process of identification and modeling of entities that plays so much part in projects of conversion of catalogs into linked open data. The help in understanding what this concept expresses within the bibliographic universe comes from object-oriented programming, which introduces the concept of modeling and management of an "object" by defining its state and behavior. But to model an object it is necessary to identify it and this process must often take place dealing with massive amounts of data, not necessarily homogeneously structured: the Entity Resolution is this set of machine processes that tries to resolve the ambiguities given by the inhomogeneity of the descriptions referable to the same entity. The adoption of these practices, in the bibliographic field, still moves the horizon of the cataloging action, which had already extended towards the more general metadating, towards that web of data that imposes a new way of understanding objects and treating them: entity modeling promises to be the third generational step in the management of bibliographic data.

KEYWORDS
Real world object; Entity; Entity resolution; Entity modeling.

Entity modeling: tracce di un percorso in evoluzione

ABSTRACT
In questo lavoro tratteremo il tema delle entità intese come oggetti reali del mondo (real-world object) e di come questo concetto sia utilizzato nell'ambito dell'entity modeling, quel processo di identificazione e modellamento delle entità che tanta parte occupa nei progetti di conversione dei cataloghi in linked open data. L'aiuto alla comprensione di cosa questo concetto esprima nell'ambito dell'universo bibliografico ci viene dalla programmazione orientata agli oggetti, che introduce il concetto di modellamento e gestione di un "oggetto" definendone uno *stato* e un *comportamento*. Ma per modellare un oggetto è necessario identificarlo e questo processo deve avvenire, spesso, trattando quantità imponenti di dati, non necessariamente omogeneamente strutturati: l'Entity Resolution è questo insieme di processi macchina che cerca di risolvere le ambiguità date dalla disomogeneità delle descrizioni riferibili alla medesima entità. L'adozione di queste pratiche, in ambito bibliografico, muove ancora l'orizzonte dell'azione catalografica, che già si era esteso verso la più generale metadatazione, verso quel web di dati che impone un nuovo modo di intendere gli oggetti e di trattarli: l'entity modeling si annuncia come il terzo passaggio generazionale nella gestione del dato bibliografico.

PAROLE CHIAVE
Real world object; Entità; Entity resolution; Entity modeling.

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

Cos'è un Real World Object

Nel suo blog Coyle's InFormation, Karen Coyle pubblica il 16 febbraio 2015 un post dal titolo Real World Objects (RWO), in risposta ad una domanda ricevuta sul significato e l'importanza del concetto di real world object[1]. Nella lista di discussione di BIBFRAME Coyle richiama il post sul RWO definendolo "this mysterious RWO thing"[2] e prova a dare una spiegazione, o meglio ad aprire una possibile riflessione sul tema. L'origine del termine e del concetto è fissato nell'ambito dell'Artificial Intelligence (AI) ed è spigato in modo chiaro: immaginiamo di dover produrre dei robot che vivano nel nostro stesso mondo, con i quali comunicare parlando di qualsiasi cosa che faccia parte del nostro mondo, sia esso un oggetto fisico o un concetto astratto. Immaginiamo, dunque, di dover stilare un elenco di cose del mondo e immaginiamo di doverlo fare in una modalità condivisibile con le macchine, e comprensibile da esse. Questa è l'ambizione fissata dall'AI e dal web semantico: poter creare e sfruttare, per dialogare con le macchine, delle raccolte (ontologie), in cui sia descritta qualsiasi cosa esistente nel mondo. Per cogliere il significato del termine real world object nell'ambito dell'AI, Coyle usa un esempio relativo al mondo bibliografico: se analizziamo un qualsiasi record bibliografico, scopriamo che i campi del record possono descrivere l'oggetto "pubblicazione", ma anche l'oggetto "record" in sé. La pubblicazione è un RWO (l'oggetto che vorrei avere tra le mani e leggere), ma anche il record è un RWO (quella descrizione testuale che io potrei richiedere ad un'altra biblioteca per migrarla nel mio catalogo). Uno dei problemi principali del record Marc in termini di capacità di veicolare un insieme di messaggi comprensibili ad una macchina, è proprio quello di non riuscire a distinguere "our metadata and the thing it describes" (Coyle 2015).
Ma come mai Caren Koyle si interroga sul real world object? Non era un termine già utilizzato in ambito bibliografico?

Il Real World Object nella disciplina catalografica

L'uso del termine real world object in ambito catalografico è particolare perché esiste e coesiste in una doppia accezione: quella ereditata dal mondo dell'AI, del web semantico e della programmazione a oggetti, e quella più tradizionalmente radicata in ambito catalografico, che utilizza il termine real world object come sinonimo di *materiale non bibliografico*, *realia*, *materiale non librario* (non-book material), per indicare oggetti tridimensionali come monete, strumenti e tessuti, posseduti dalle biblioteche ma che non rientrano nelle categorie tradizionali del materiale librario. Contestualmente a questo significato, però, il termine real world object viene ufficialmente e definitivamente assorbito nel gergo biblioteconomico: nel 2015 viene costituito il PCC Task Group on URIs in MARC (URI TG) con l'obiettivo di identificare e indirizzare le scelte per l'arricchimento del record MARC con identificatori (URI). Il lavoro del gruppo pone le basi per facilitare la transizione dei dati MARC verso i linked data: tra le diverse proposte e soluzioni quella di includere

[1] La sintassi utilizzata per questo termine non è omogenea: Real-World Object, real-world object e real world object sono le diverse formulazione sintattiche più frequentemente utilizzate nelle diverse fonti consultati.
[2] <LISTSERV 16.0 - BIBFRAME Archives (loc.gov)>.

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

un nuovo sottocampo $1 per la registrazione dell'URI del RWO[3] nei campi relativi a nomi, titoli, soggetti, classificazioni etc. Nel 2017 è pubblicato il documento MARC Proposal no. 2017-08 dal titolo: Use of Subfield $0 and $1 to Capture Uniform Resource Identifiers (URIs) in the MARC 21 Format. Scopo della proposta è quella di delineare un criterio per arricchire i record bibliografici e di authority con gli identificatori URI in un modo che siano chiaramente distinguibili:

- URI che identificano un record che descrive un oggetto (URI come puntatore ad una descrizione);
- URI che identificano l'oggetto stesso.

A tal fine, il documento propone di limitare l'uso del sottocampo $0 agli URI e ai numeri di controllo che si riferiscono *al record che descrive una cosa* e di definire un nuovo sottocampo, $1, per gli URI che si riferiscono direttamente al *Real World Object*. La proposta diventa operativa e l'Appendix A - Control Subfields del MARC 21 Bibliographic Format ufficialmente introduce questa importante distinzione tra identificatori per descrizioni e identificatori per RWO**.**

Questa distinzione, in qualche modo, chiude il cerchio rispondendo a quella perplessità chiaramente espressa da Karen Coyle, ma comune a molti esperti di metadatazione e di modellazione dei dati, rispetto alla mancata distinzione, nel record bibliografico e di authority, tra "our metadata and the thing it describes".

Real World Object e Real World Entity

A complicare l'intero quadro terminologico si aggiunge l'uso del termine Real World Entity spesso in alternativa o in concomitanza con il termine Real World Object. Il Real World Entity è definito su alcuni siti come un'entità con una posizione fisica all'interno dell'universo, quindi con una definizione che, in sé, non lascia intendere nulla di diverso rispetto al Real World Object: *object* e *entity* sono usati come alternativi l'uno all'altro. E lo stesso accade in molta della letteratura che richiami queste tematiche, in cui i termini *real world object* e *real world entity* sembrano rimandare allo stesso concetto. Tuttavia, nonostante il nuovo rimescolamento di carte che l'uso di questa terminologia, nei termini descritti, sembra produrre, la traccia da seguire per distinguere la sottilissima linea di demarcazione semantica tra questi termini sembra essere stata individuata: sappiamo che viviamo in un mondo costituito da entità di vario tipo ("cose" fisiche e concettuali), sappiamo che spesso queste entità sono estremamente complesse e sappiamo che nessun sistema che voglia rappresentarle potrà avere l'ambizione di esprimerle nella loro interezza: ci sarà sempre un aspetto di un'entità che non sia facile cogliere o che il contesto per il quale quella "cosa" sia descritta non richieda di essere rappresentata. L'ipotesi che stiamo provando a verificare è che nelle espressioni *real world entity* e *real world object* ci possa essere la stessa relazione semantica che possiamo rilevare nel binomio *entità* e *identità*.

[3] Per una chiara e completa disamina sui lavori del gruppo PCC Task Group on URIs in MARC si rimanda all'articolo di Jackie Shieh, dal titolo PCC's Work on URIs in MARC, pubblicato nel 2019 in Cataloging & Classification Quarterly (Shieh 2020).

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

Il Real World Object nell'Object Oriented Programming

La programmazione orientata agli oggetti, o più semplicemente programmazione ad oggetti, si basa sulla definizione di *classi* che contengono la dichiarazione delle strutture dati e le procedure che operano su di esse. La classe costituisce un modello o un progetto per quelli che poi saranno gli *oggetti* che deriverò da essa. Una sorta di template che definisce la forma dell'oggetto, utilizzato per poi modellare i singoli oggetti. Una classe definisce qualcosa in termini di:

- *stato*: le variabili che costituiscono quel tipo di cosa
- *comportamento*: i comportamenti (metodi, procedure) che ha quel tipo di cosa.

Questo modello definito dalla classe viene poi istanziato nei diversi oggetti, istanze, create a partire dalla medesima classe. Nel contesto dell'OOP un oggetto software viene creato a partire (e come rappresentazione di) un oggetto reale del mondo, un real world object: il programmatore è chiamato a riprodurre in una dimensione diversa (quella del software, appunto) il mondo reale, con i suoi tanti oggetti (siano essi fisici o concettuali). E così come gli oggetti del mondo reale, quindi i real world object, hanno tutti uno *stato* (degli attributi che li identifichino) e un *comportamento* (la capacità di fare qualcosa) allo stesso modo, gli oggetti software devono poter essere modellati con uno stato e un comportamento. Ma perché rappresenta un real world object e non un real world entity? La programmazione orientata agli oggetti dichiara, in modo piuttosto esplicito, la difficoltà o addirittura l'impossibilità a rappresentare in un oggetto software la complessità di un'entità: l'oggetto software è la rappresentazione di un real world object e non di un real world entity perché l'entità, nella sua ricchezza, non potrà mai essere rappresentata in modo esaustivo. L'OOP costruisce oggetti che rappresentino una particolare "faccia" dell'entità, e solo quella, adatta e funzionale allo specifico contesto in cui venga utilizzata. Questo è uno dei pilastri dell'OOP, chiamato *astrazione*: l'informatica non ha l'ambizione di rappresentare tutte insieme le mille caratteristiche di una entità, ma *seleziona* quelle adatte al contesto in cui quell'oggetto debba essere calato (nascondendo o ignorando del tutto le tante altre caratteristiche - di stato e comportamento - ascrivibili a quella medesima entità). Questo principio è collegato con un altro dei pilastri dell'OOP: il *polimorfismo*, che è definito come la capacità di usare lo stesso nome per fare differenti cose, oppure, più chiaramente, la capacità di rappresentare *molte forme della singola entità*.

L'Entity Resolution e l'Entity Modeling

Il meccanismo del "creare" un oggetto, mettendolo in relazione con altro, così tipico di questo paradigma di programmazione, non è affatto diverso da quel nuovo modo di intendere l'attività del catalogatore nell'universo bibliografico, sempre più orientato alla identificazione delle entità che partecipano a questo universo, e alla loro "modellazione". Modellare un oggetto significa, dunque, individuare quei tratti salienti che lo rendono ciò che è, e che gli consentono di essere o fare qualcosa nel mondo, o per lo meno in un determinato contesto. Ma se concordiamo con l'immagine di un'entità complessa che sia presente in forme diverse sulla medesima fonte informativa o in fonti differenti, possiamo immaginare quanto complesso sia il meccanismo dell'identificazione dell'entità. Rinunciare all'ambizione di descrivere nella sua interezza la complessità dell'entità non significa rinunciare anche all'ambizione di *identificarla* nelle sue tante espressioni: l'informatica

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

accetta il limite di non poter rappresentare in un unico oggetto la complessa personalità di Lewis Carroll, e probabilmente costruirà due o più "oggetti" (uno come Charles Lutwidge Dodgson, autore dell'opera *The game of logic* (1887) e uno come Lewis Carroll, autore del celebre *Alice's Adventures in Wonderland*). Ma in una realtà più complessa, come quella del web semantico e dell'intelligenza artificiale, possiamo rinunciare all'idea di riconoscere dietro questi due "profili" la medesima entità? Questo è esattamente il compito dell'Entity Resolution (abbreviato ER), quello di identificare tutte le menzioni che rappresentano la stessa entità all'interno della stessa base di conoscenza oppure in basi di conoscenza multiple (Zhu et al. 2016).

L'Entity Resolution (*risoluzione delle entità*) è il processo che risolve le entità e rileva le relazioni utili a identificarle. Il processo nel suo insieme genera un database di entità, in cui, dunque, i dati che identifichino una entità sono registrati per consumo (ricerche, statistiche etc.) o anche per iterare i processi di "entificazione" e renderli sempre più efficaci.

Il punto di partenza, dunque, per i processi di Entity Resolution, è costituito da un insieme di informazioni che esprimono l'identità di un particolare real world entity: questa unità informativa, nell'ambito dell'ER, viene spesso definita *profilo* e a noi ricorda molto il risultato di quella *astrazione* che abbiamo trovato come pilastro dell'OOP.

In un contesto informativo sempre più ampio, come quello che il web semantico propone, in cui le basi di dati sono eterogenee e non necessariamente autorevoli, i profili sono spesso sporchi, incompleti, incorretti o ridondanti: il successo dell'ER è quello di riuscire a identificare e integrare profili anche molto diversi ma che rimandino, in realtà, alla medesima entità.

Se definiamo due profili relativi alla stessa entità, presenti nella medesima fonte dati, come "duplicati", allora possiamo dire che una fonte dati pulita è quella in cui non esistano duplicati, mentre una fonte dati è sporca quando esistano profili duplicati. In realtà la classificazione di una fonte è ben più complessa di questa esemplificazione: nella stessa fonte di dati si possono trovare casi di entità molto ben modellate e altre con molteplici profili. La "veracità" di una fonte è un parametro complesso che si basa su numerosi algoritmi, e che affida una percentuale piuttosto alta di successo alla qualità del dato di origine.

L'obiettivo dei processi che puntano alla veracità del dato è quello di *collegare* profili o descrizioni di entità diverse *(record linkage)* e deduplicare, dunque individuare i duplicati e risolverli.

I complessi processi di ER raramente si affidano a singoli dati, o a profili poveri: più il profilo utilizzato è ricco di caratteristiche, più attributi sono espressi a definire una particolare identità dell'oggetto descritto, più possibilità si avranno di identificarlo. Ma perché una macchina possa identificare un oggetto, deve prima di tutto conoscerne i contorni: deve sapere cosa cercare e come si aspetta di trovare quell'oggetto. Deve sapere, dunque, come quella cosa sia *modellata*. La costruzione dell'oggetto, la definizione del profilo in una modalità che sia rispondente alle esigenze prima di tutto di identificazione, è un'attività della massima importanza, soprattutto in un contesto aperto e tremendamente sconfinato come quello del web. E per intendere cosa sia l'entity modeling, ripartiamo da quell'oggetto che l'OOP crea attribuendo uno *stato* e un *comportamento*, e che il web semantico riformula, arricchisce in modo da renderlo non solo usabile, ma anche comprensibile alle macchine. Il meccanismo di creazione di un oggetto nell'OOP parte dalla definizione di un modello, quanto abbiamo definito come una sorta di *matrice* o template e che serve per creare tante istanze che da quel modello prendono, appunto, la forma. Queste matrici (le classi, con le loro caratteristiche di stato e comportamento) possono essere costruite specifica-

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

tamente per il singolo oggetto, nell'ambito di uno specifico progetto (il che rende quegli oggetti poco "usabili" al di fuori del contesto originario) oppure possono essere definite a livello globale, come risultato di un accordo che una comunità decida di sottoscrivere e condividere. Le ontologie e i data model, in tutti i contesti in cui siano utilizzati, sono esattamente questo: una matrice che definisce la forma che avranno le rappresentazioni di un real world object nella maniera che sia il più possibile fedele all'entità rappresentata, in accordo con quanto definito da una comunità.

I modelli e le ontologie dell'universo bibliografico

Non potendo rappresentare l'entità nella sua interezza, abbiamo concordato di scegliere un criterio di *utilità* per filtrarne o selezionarne i caratteri utili ad un determinato fine: per quali scopi rappresento questa entità? Nel contesto della disciplina catalografica la definizione dei principi e delle regole catalografiche, la scelta delle intestazioni (poi diventati access point), la definizione delle entità da rappresentare, sono tutti fattori strettamente collegati agli obiettivi e funzioni del catalogo che, già a partire da Cutter, quindi già alla fine del XIX secolo, sono stati fortemente influenzati dalle *esigenze dell'utente (user's tasks)* Nella sua opera *Rules for a Printed Dictionary Catalogue* Cutter definisce gli obiettivi del catalogo proprio in funzione dell'utilità per gli utenti (Cutter 1876). Questo criterio delle esigenze dell'utente ha guidato tutte le successive revisioni dei modelli e delle regole catalografiche nonché la definizione delle entità necessarie a supportare tali bisogni.
L'input a definire principi, regole, modelli e infine ontologie viene, dunque, prima di tutto da quel che gli utenti si aspettano di vivere nella loro esperienza di utilizzo del catalogo. Per lo meno in ambito bibliografico chi si pone di fronte al compito di definire dei modelli catalografici, di modellare la conoscenza, lo fa avendo ben chiaro in mente quali siano gli obiettivi di ciascun elemento che entri a far parte di quel modello.

Un esempio di entity modeling: l'Opus in Share-VDE

L'iniziativa Share-VDE (Share Virtual Discovery Environment),[4] che riunisce, arricchisce e integra i cataloghi bibliografici e di authority di una vasta comunità di biblioteche in un ambiente condiviso basato su linked data, ha al proprio interno diversi gruppi di lavoro,[5] che affrontano alcuni ambiti più complessi o per i quali un'analisi ulteriore e specializzata sia particolarmente richiesta. I gruppi sono coordinati da un *Advisory Council* che traccia il percorso dell'iniziativa e coordina le diverse anime che vi partecipano, nonché i rapporti e le relazioni con iniziative e gruppi affini.
Il gruppo *Sapientia Entity Identification Working Group - SEIWG*, che si occupa di tutti i temi relativi al modellamento dei dati, ha lavorato alla revisione del modello dati di Share per renderlo più vicino alle esigenze di ricercabilità e identificazione delle risorse, soprattutto in un contesto così esteso e ricco come quello prodotto dall'iniziativa. Il percorso di analisi e studio ha avuto come risultato l'identificazione di una nuova entità, l'*Opus*, non già presente nel modello BIBFRAME

[4] <https://wiki.share-vde.org/wiki/Main_Page>.
[5] <https://wiki.share-vde.org/wiki/ShareVDE:Members/Share-VDE_working_groups#SVDE_Advisory_Council_.28AC.29>.

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

che la comunità di Share adotta come ontologia principale per la resa dei propri cataloghi in linked open data. Il percorso di analisi e di proposte e ipotesi di soluzione è stato piuttosto lungo e ha visto il coinvolgimento di molte biblioteche nordamericane e europee nella discussione intorno alla possibile estensione del modello BIBFRAME.

Il punto di partenza: l'ontologia BIBFRAME

Share-VDE è un'iniziativa che nasce con lo scopo di accompagnare e supportare le biblioteche in questo delicato momento di transizione dal record Marc ai modelli basati sulle entità, con particolare attenzione a BIBFRAME, che è stato adottato, appunto, come ontologia centrale nei processi di conversione dei dati in linked open data. La presenza di tante biblioteche americane, a partire dalla Library of Congress, ha orientato fortemente la scelta di adozione di BIBFRAME, anche se l'iniziativa partiva dalle suggestioni offerte da un progetto tutto italiano, quello delle biblioteche universitarie del sud d'Italia, chiamato Share Catalogue, che aveva già operato questa scelta molto determinata di adozione di BIBFRAME. D'altra parte, l'ingresso nella comunità di Share-VDE della National Library of Norway, della National Library of Finland e della British Library, tutte biblioteche con un profondo legame con il contesto culturale europeo e con l'IFLA, ha suggerito una importante riflessione in merito alla compatibilità del modello BIBFRAME con quello definito dall'IFLA come armonizzazione della famiglia FR, quindi con l'IFLA Library Reference Model (abbreviato IFLA LRM).
Fino al giugno 2021, con l'introduzione ufficiale dell'entità Hub, il modello BIBFRAME nella versione 2.0 organizzava le informazioni utili a modellare l'universo bibliografico in tre livelli principali di astrazione:

- **Work**: contenuto intellettuale di una risorsa.
- **Instance**: una singola manifestazione materiale di un Work.
- **Item:** il singolo esempio di un'Istanza.

Queste entità principali sono corredate di una serie corposa di entità relazionate (tra cui, ovviamente, quella degli Agenti).
Di contro, il modello IFLA LRM eredita i livelli delle varie declinazioni della famiglia FR e sviluppa un modello concettuale ben più complesso, che comunque ripropone come entità principali quelle di FRBR:

- **Work**: il contenuto intellettuale o artistico di una determinata creazione.
- **Expression**: una determinata combinazione di segni che veicola un contenuto intellettuale o artistico.
- **Manifestation**: un insieme di tutti i supporti che si presume condividano le stesse caratteristiche per quanto riguarda contenuto intellettuale o artistico e aspetti di forma fisica.
- **Item**: un oggetto o oggetti che recano segni destinati a veicolare un contenuto intellettuale o artistico.

La mancanza di un elemento apicale in BIBFRAME che indirizzi l'utente su una possibile opera originale, quell'opera creativa che era nella mente del suo autore e che si è poi espressa in molti modi differenti, non può essere risolta con la proprietà "translation of" di BIBFRAME perché questa proprietà può indistintamente relazionare un'opera originale e la sua traduzione così come

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

una traduzione in una lingua ma a partire da un'altra espressione non originale e dunque da un'altra traduzione. Il che pone tutti i Work di BIBFRAME sulla medesima linea orizzontale, senza alcuna indicazione di un'opera originaria.
Ma il problema di un'opera con una storia editoriale molto ricca in cui si perda traccia del punto di partenza da cui il reticolo espressivo ha poi preso forma si sente molto e non è un caso che IFLA LRM introduca il concetto di *Espressione rappresentativa* che dovrebbe essere, qualora identificabile, la prima espressione di un'opera, dunque quella che meglio esprima l'opera originale, quella più vicina a quanto idealmente pensato dal suo creatore.
Il risultato delle prime riflessioni del gruppo SEI su questo tema e i tanti confronti con esperti di dominio, conduce, nel gennaio 2019, ad una prima ipotesi di modellamento di un'entità, inizialmente definita *SuperWork*, che potesse raccogliere sotto di sé tutte le diverse espressioni relazionate, qualcosa, dunque, molto vicino al concetto di Work di FRBR e di IFLA LRM. Ma l'introduzione di una nuova entità come estensione, locale o ufficiale, di un modello riconosciuto da una comunità - BIBFRAME - potrebbe generare, prima di tutto, un problema di interoperabilità. Problema sentito fortemente anche da tutte le biblioteche interessate a modellare i propri dati secondo l'ontologia BIBFRAME ma non disposte a perdere la possibilità di un dialogo e uno scambio agile di dati e servizi con l'altrettanto riconosciuta comunità riferentesi al modello IFLA LRM.
Intanto, la Library of Congress, anche seguendo le suggestioni date dalle conversazioni e i confronti tenuti nel gruppo di lavoro del SEI, proponeva, prima solo in test (nel giugno 2019) e poi in ambiente di produzione (McCallum 2022), quell'elemento "aggregante" costituito dall'Hub, definito come *una risorsa astratta che funziona come collegamento tra due Work.*
A questo punto gli elementi per una comparazione puntuale tra i tre modelli ontologici, quello di Share-VDE con le proposte modifiche al modello originale, quello di BIBFRAME e quello di IFLA LRM erano maturati, dando così il via ad un'attenta valutazione comparativa ai fini, appunto, di garantire una più efficace risposta al bisogno informativo dell'utente ma anche una piena compatibilità con i due maggiori e più diffusi modelli bibliografici. Nel gennaio 2020, anche per sancire la funzione semantica attribuita al livello apicale definito SuperWork e per evitare confusioni con altre definizioni passate di "superwork", il gruppo del SEI rinomina questa nuova classe come Share-VDE Opus (svde:Opus).

La definizione della classe svde:Opus e il suo modellamento

Definita a livello teorico la classe svde:Opus, bisognava renderla operativa, capace, cioè, di essere modellata con proprietà e relazioni, in modo da renderla efficace da un punto di vista di finalità ma anche utilizzabile da altre comunità. Il gruppo di lavoro SEI ha ipotizzato, così, tre possibili scenari di modellamento dell'Opus, cui poi se ne è aggiunto un quarto come variante dello scenario terzo: l'esercizio di modellamento è servito al gruppo per valutare tutti i pro e i contro di ciascuno scenario, partendo da alcuni elementi chiave intesi come "desiderata" e senza mai perdere di vista il principio dell'interoperabilità nella più ampia comunità del web. Negli schemi di rappresentazione dei 3 scenari di partenza, che sono quelli originali prodotti e discussi nell'ambito del gruppo di lavoro SEI, la formulazione della nuova entità riporta ancora la nomenclatura SuperWork invede di Opus:

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

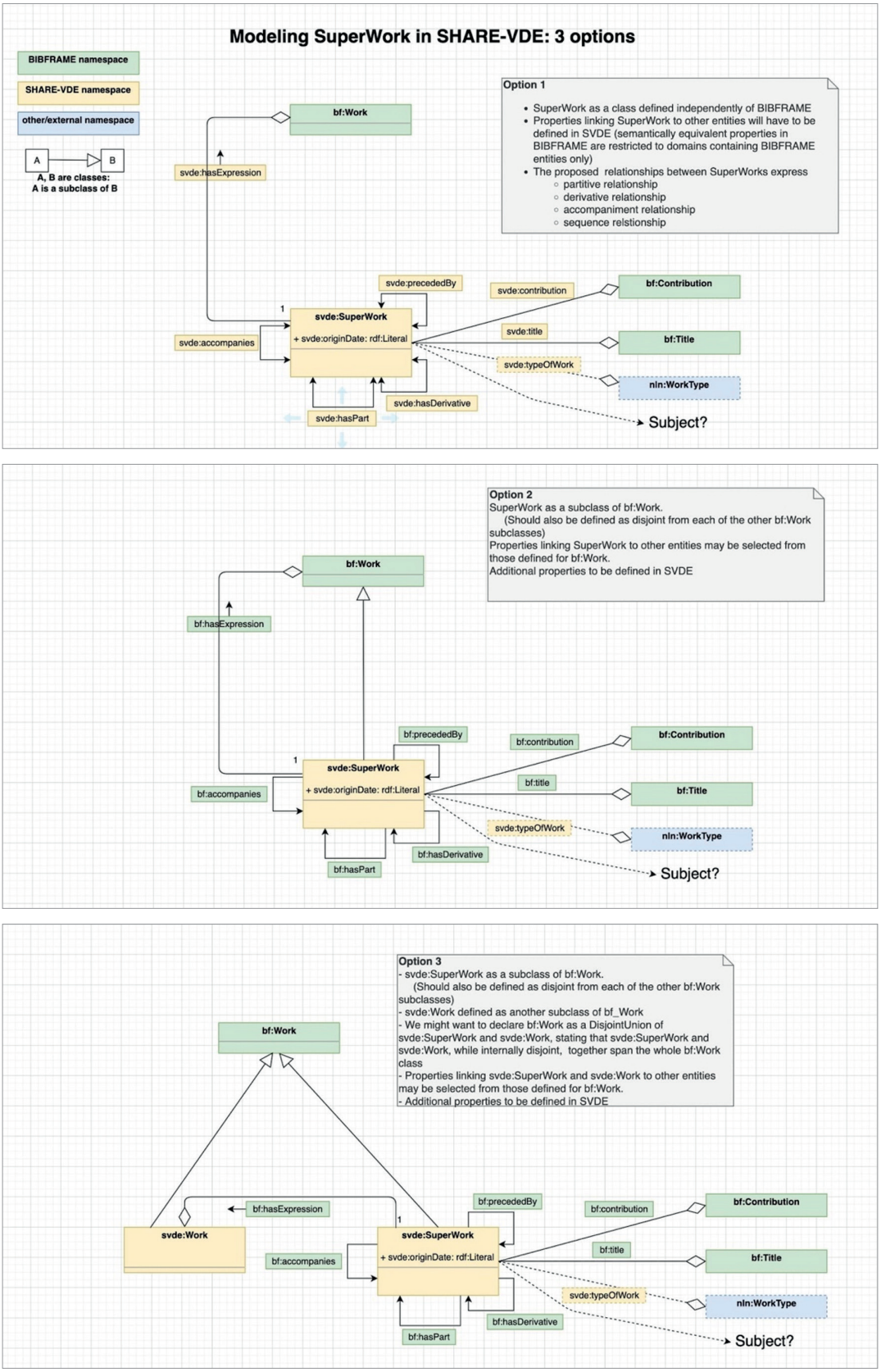


Figure 1-3. I tre scenari di modeling ipotizzati per l'Opus.

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

Un leggero rimodellamento dell'opzione 3, quella verso la quale il gruppo del SEI sempre più si orientava, è stato proposto e formalizzato in quella poi definita opzione 4, in cui, per altro, appare la formulazione definitiva di Opus:

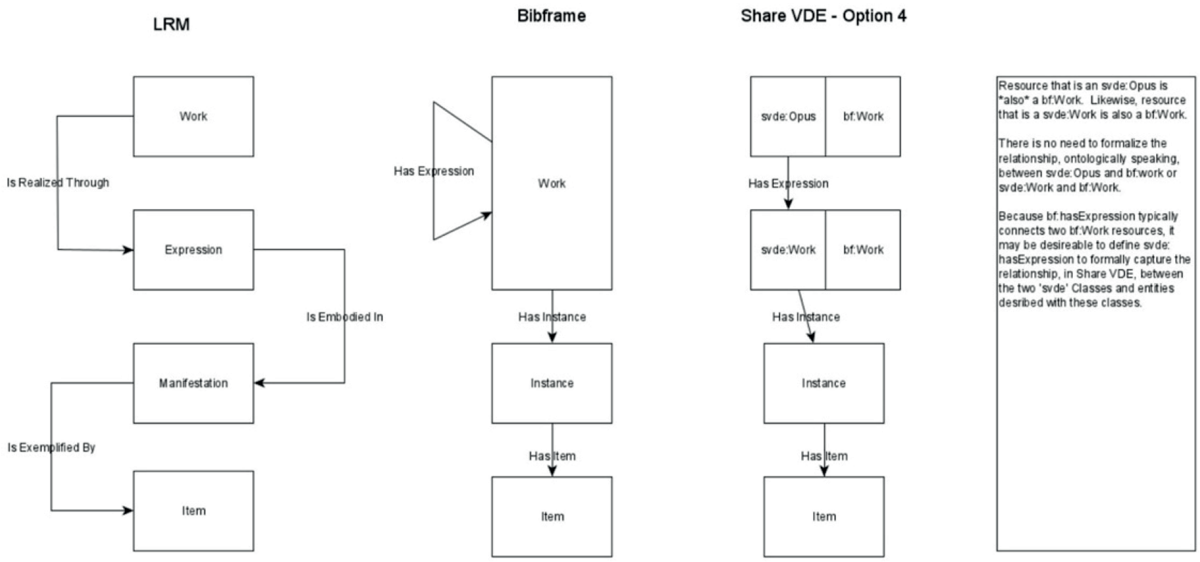


Figura 4. La quarta opzione di modeling dell'entità Opus.

La discussione del gruppo di lavoro si è così concentrata sull'analisi comparativa del modello 3 e della sua versione 4, con il preciso mandato dell'Advisory Council di studiare e formalizzare lo scenario che meglio si adattasse allo scopo di identificazione, descrizione, conversione e mantenimento dei dati delle biblioteche, tenendo conto anche delle sfide dell'interoperabilità con altri modelli. Per arrivare ad una decisione congiunta, dunque, il gruppo SEI ha formulato un elenco di "desiderata" per il modellamento dell'entità, con le rispettive formalizzazioni realizzabili nelle due opzioni selezionate come possibili, elaborando a conclusione dello studio comparativo una tabella di riepilogo per pesare i pro e i contro dei due scenari di modeling, che porterà alla decisione finale, quella di adottare lo scenario nella sua versione 4:

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

Opzioni del modello	Pro	Contro
Opzione 4	1. Soddisfa la caratteristica del modello desiderata in modo flessibile 2. Concede tempo affinché le migliori pratiche si sviluppino ulteriormente e forniscano miglioramenti agli algoritmi di istanziazione 3. Nonostante le mappature dei cluster di entità e gli algoritmi di istanziazione siano diversi, l'approccio è simile a quello della Library of Congress rispetto al bf:Hub	1. Senza una definizione formale delle sottoclassi, la documentazione delle mappature dei cluster di entità e proprietà diviene imprescindibile 2. (Forse non è un "contro") Con lo sviluppo delle buone pratiche e degli standard, l'applicazione delle caratteristiche del modello qui presentato può essere auspicabile a livello di ontologia BIBFRAME
Opzione 3	1. Soddisfa la caratteristica del modello desiderata (alcune perplessità sono annotate nella colonna "Contro") 2. Fornisce una definizione formale di svde:Opus e svde:Work come sottoclassi	1. L'uso di sottoclassi pone l'interrogativo su come preservare l'equivalenza tra svde:Work come sottoclasse di bf:Work e bf:Work in Share-VDE. Ciò rivela questioni di interoperabilità circa l'input e l'output verso/da fonti esterne. 2. Diverse domande potrebbero beneficiare di ulteriori analisi e sviluppo di buone pratiche.

Figura 5. Schematizzazione dei pro e dei contro degli scenari 3 e 4.

Le valutazioni sulla scelta dell'opzione 4 sono condivise con la comunità di Share e con tutti coloro che hanno contribuito alla riflessione su un così delicato intervento di modeling.

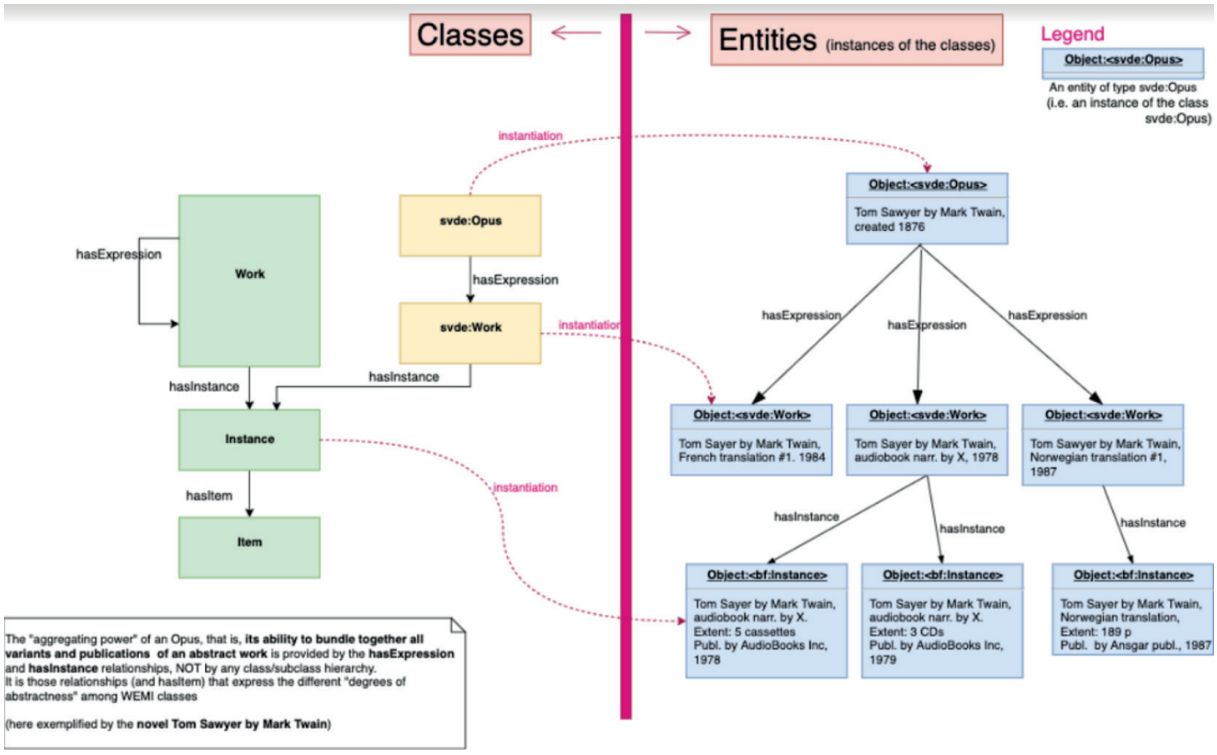


Figura 6. L'Opus e le sue relazioni qui esemplificate nell'istanziazione del romanzo Tom Sawyer di Mark Twain.

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

In questa soluzione di modellamento la decisione centrale, quella che rende davvero compatibile questa estensione ontologica con BIBFRAME e con IFLA LRM (e RDA) è la seguente: la risorsa di tipo svde:Opus è anche un bf:Work (quindi è *un tipo di* bf:Work) così come la risorsa svde:Work è anche un bf:Work. Le due entità modellate per Share-VDE diventano *tipi di* Work BIBFRAME, essendo così perfettamente compatibili sia con BIBFRAME che con le entità Work ed Expression di IFLA LRM.

L'Entity modeling come terza generazione della disciplina catalografica

Quanto fin qui detto dell'Entity Resolution e dell'Entity modeling riguarda soprattutto elaborazioni massive di dati, quindi processi macchina attivati in progetti di conversione da un formato all'altro (per esempio da Marc a RDF) o nell'ambito del machine learning. Fino a pochissimo tempo fa. La logica *entity-oriented* sta maturando e coinvolgendo in modo sempre più evidente le pratiche manuali di catalogazione: come gli ILS e le piattaforme di gestione dei dati si posizioneranno rispetto all'avvicinamento sempre più prossimo della biblioteconomia ai linguaggi e alle tecniche del web, è un quesito aperto, che comincia a dare risposte pratiche sul fronte delle conversioni e sul fronte dei sistemi di discovery. Ma il cambiamento di orizzonte prospettato dal web semantico sta superando il contesto tradizionale delle conversioni e dei discovery e sta avviando ipotesi su nuovi scenari operativi, con definizione di casi d'uso focalizzati sulla creazione nativa di dati in rdf, l'analisi dell'impatto sui processi produttivi, lo sviluppo di moduli di catalogazione in linked data. Solo a titolo esemplificativo si citano:

- il progetto di sviluppo di un modulo di catalogazione in rdf, Libris (XL), voluto già nel 2018 dalla National Library of Sweden per gestire la nuova versione dello Union Catalogue nazionale (Wennerlund e Berggren 2019);
- l'iniziativa LD4P - Linked Data for Production, che nella Fase 2 si è concentrata sullo sviluppo di un ambiente di catalogazione basato su BIBFRAME e chiamato Sinopia (Schreur 2019);
- l'editor BIBFRAME della Library of Congress, un tool pensato per consentire di modellare le descrizioni bibliografiche secondo l'ontologia BIBFRAME.

Guardando questi editor e la loro focalizzazione sull'entità invece che su un record, si capisce a pieno il cambio di prospettiva e di orizzonte che anche nelle biblioteche sta maturando: quando un catalogatore comincia a ragionare su *come modellare un oggetto* (per esempio un libro, oppure una persona, oppure un evento), quali attributi e quali relazioni aggiungere per meglio rappresentare l'oggetto che sta descrivendo, per renderlo il più identificabile possibile anche al di fuori della biblioteca, allora ha già smesso di catalogare: sta *modellando l'entità*. Questo tema si riallaccia ed estende il tema del cammino della catalogazione verso la metadatazione, che è un processo concretamente già da molto tempo avvenuto (Gorman 2018, 121), e che ha ancora recentemente aperto un vivace dibattito teorico in Italia (Guerrini 2020) (Guerrini 2022). Pensando a cosa significhi l'entity modeling e a quanto incarni quell'auspicato passaggio dal record al real world object, mi piacerebbe provare a indicare un altro modo di guardare a questa evoluzione, come ad una sorta di *cambiamento di visione prospettica*, quasi come se il processo fosse quello di guardare la stessa cosa ma da un'altra visuale. Riepiloghiamo gli elementi chiave di ciascuno di questi "scenari",

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

quello della catalogazione, della metadatazione e dell'entity modeling, per provare a definire i confini di ciascuna di queste attività.

a) *Catalogazione*: questo termine esprime bene la dimensione più tradizionale dell'attività bibliografica, quella di rappresentare in un catalogo il posseduto di una biblioteca o altre informazioni bibliografiche, attraverso l'applicazione di regole e standard che abbiano come motrice l'interesse dell'utente. La dimensione "spaziale" sufficiente per garantire questa dinamica di relazione tra l'attività della biblioteca e i suoi utenti è quella definita dalla biblioteca stessa e in qualche modo da essa delimitata: il catalogo rispecchia, per lo più, quanto posseduto da una biblioteca o quanto veicolato dalla biblioteca; l'utenza è quella della biblioteca stessa.
b) *Metadatazione:* la definizione più semplice di questo termine è quella di "dati sui dati" o "informazioni sulle informazioni" (Riley 2004). Nei diversi contesti di utilizzo, il termine è stato usato per indicare le informazioni relative a qualsiasi *cosa*: libri, oggetti museali, finding aids per materiale archivistico, immagini: "*Broadly speaking, metadata encapsulate the information that describes any information-bearing entity*". Come Zeng e Qin ricordano nella loro fondamentale opera *Metadata*, dagli albori dei cataloghi e degli indici scritti a mano e stampati fino ai giorni nostri, quelli dei servizi web e delle app, la natura e l'obiettivo di descrivere le entità portatrici di informazioni sono rimaste più o meno invariate (Zeng e Qin 2016). Tuttavia, i metodi e le tecnologie sono cambiati in modo significativo. L'estensione dell'orizzonte di riferimento della biblioteca e degli istituti della cultura in generale, da una dimensione locale ad una dimensione sempre più ampia, l'arricchirsi dei materiali entrati a far parte delle collezioni o anche solo ad essi referenziati, l'allargamento dei servizi proposti dall'aumento esponenziale del digitale, con anche il moltiplicarsi delle relative mansioni di gestione, sono tutti fattori che hanno portato ad una vera esplosione di metadati, al punto che il termine "metadatazione" sta soppiantando lo stesso termine "catalogazione".
c) *Entity modeling*: abbiamo parlato dell'entity modeling, di come realizzi quel cambiamento di mentalità prima che tecnologico che pone l'entità, o il real world object, al centro della propria attenzione, come oggetto da costruire e modellare. Il panorama di riferimento non è neanche più quello del web tradizionale ma quello del web semantico, il web dei dati, in cui gli oggetti sono costruiti seguendo il paradigma dei linked data e aggiungendo "semantica" alle informazioni, sì da renderle condivisibili con le macchine. La tecnologia, gli standard, i protocolli apparentemente si complicano, ed in certa misura questo davvero accade. Ma si tratta, prima e soprattutto, di un cambio di ottica, un nuovo passo verso un universo più esteso, quello globale del web, dove anche la terminologia deve superare il proprio limite di riferimento a singoli domini.

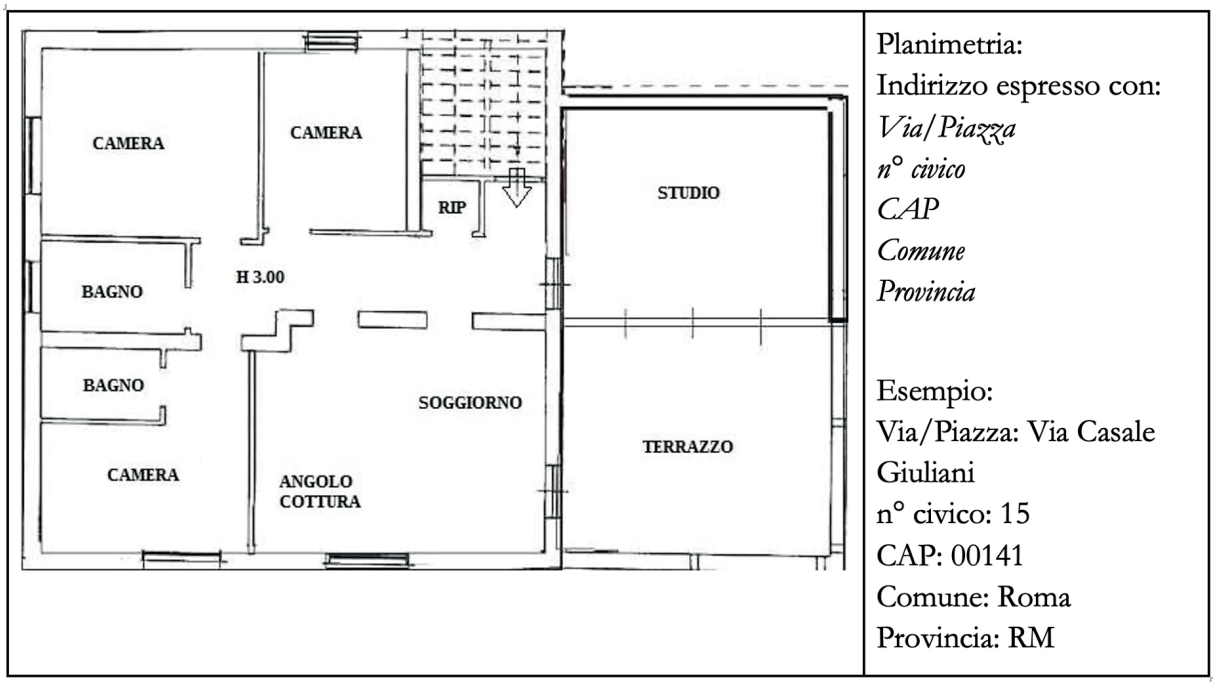
Se provassimo ad analizzare il risultato concreto delle tre operazioni elencate sopra, catalogazione, metadatazione, entity modeling, ci accorgeremmo che non c'è alcuna differenza nel risultato pratico: un record bibliografico, un set di metadati, un dataset di dati rdf. Metadati, di differenti tipologie e livelli di profondità, ma comunque metadati.

Ma se invece quei tre scenari provassimo ad analizzarli in modo diverso, cambiando la nostra ottica di visuale, ci accorgeremmo di quanto siano differenti e di quanto sia legittimo un cambio anche radicale di terminologia. Di seguito si rappresenta un oggetto (un appartamento) visto in tre dimensioni "spaziali" diverse:

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

- la dimensione locale, con una *planimetria* che non ha riferimenti al di fuori di sé e risponde all'esigenza di rappresentare l'oggetto senza alcun riferimento chiave al contesto esterno. I metadati associati alla planimetria sono quelli tipici per identificare uno stabile in una data località: un indirizzo con i suoi campi. Questa dimensione rappresenta la catalogazione, e la sua funzione di rappresentazione di una realtà più locale;
- la dimensione estesa al municipio, con una *mappa catastale* in cui quello stesso appartamento non è più rappresentato nel suo "isolamento", ma rispetto ad un catasto urbano ben più ampio. Lo stesso oggetto, lo stesso appartamento, viene identificato rispetto ad un orizzonte più ampio, e gli elementi descrittivi necessari a identificarlo sono diversi e soprattutto sono in relazione ad altri oggetti, e per altre finalità: sezione urbana, foglio, particella, subalterno. Questa dimensione rappresenta invece quella della metadatazione, e la sua funzione di rappresentare dati più ricchi e che abbracciano funzioni diverse: metadati descrittivi, metadati amministrativi (tecnici, di conservazione, sui diritti), metadati strutturali e linguaggi di marcatura;
- la dimensione geospaziale, con sempre lo stesso oggetto visto da una dimensione ancora più ampia: in questa dimensione i punti di riferimento per l'identificazione diventano quelli di latitudine e longitudine, dunque con riferimento al più vasto "globo". Questa dimensione è quella dell'entity modeling, il cui retroterra di realizzazione è quello ben più ampio del web (semantico).



Planimetria:
Indirizzo espresso con:
Via/Piazza
n° civico
CAP
Comune
Provincia

Esempio:
Via/Piazza: Via Casale Giuliani
n° civico: 15
CAP: 00141
Comune: Roma
Provincia: RM

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

Figura 7-9. Rappresentazione di un medesimo oggetto da tre prospettive visuali e funzionali diverse.

Importante è sottolineare che il cambio di visuale non è fine a sé stesso, non è una scelta filosofica né tanto meno una moda, ma ha una sua necessità funzionale: lo stesso dato viene identificato, descritto, comunicato per finalità differenti. Questo potrebbe sembrare, per una biblioteca, un allontanamento dall'obiettivo principale o tradizionale, quello del proprio catalogo e della propria utenza. In realtà è l'espressione di un positivo e direi inevitabile allargamento dell'orizzonte: è come se le cose si guardassero da una visuale sempre più ampia, in un orizzonte sempre più esteso, partendo dalla dimensione locale, del proprio catalogo, passando per l'inclusione di nuovo materiale, di nuovi servizi, di nuovi tipi di utenza, di contaminazioni sempre più evidenti con mondi "affini" (cross-domain e interoperabilità) e arrivando, per ora, al web, in cui ogni "cosa" locale di-

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

venta tremendamente piccola e limitata se confrontata con la vastità che esso, il web, rappresenta. Ed è inevitabile che un cambio di orizzonte generi anche un cambio di linguaggio: se prima era sufficiente parlare di *documento*, pur con tutti i dubbi su cosa davvero sia un *documento*, ora bisogna adottare termini più globalmente riconoscibili, come quello di *risorsa*, che nella sua generalità e neutralità meglio esprime la totalità delle cose descrivibili nel web.

Conclusioni

Le riflessioni proposte in questo studio partono da un concetto ormai quasi abusato quando si parla di linked data applicati all'ambito bibliografico, e cioè il passaggio dal record al real world object, per provare a definire meglio cosa si intenda per real world object e come questo termine sia entrato nel gergo bibliografico e sia utilizzato. L'indagine qui svolta è strettamente collegata ad un'analisi ancora in corso sul binomio tre *entità e identità*, binomio indagato sotto un profilo soprattutto filosofico, per arrivare, ovviamente, a individuare un criterio di applicazione di questa terminologia nell'ambito della disciplina catalografica: la distinzione tra real world entity e real world object, o più semplicemente tra entità e real world object, rimanda alla proposizione di un'entità complessa che si esprime attraverso diverse identità. L'informatica, soprattutto nella sua declinazione di programmazione orientata agli oggetti, ci aiuta a meglio focalizzare questa ipotesi di modellamento dell'universo bibliografico, che già da tempo si era appoggiato ai modelli entità-relazione ma che fatica a focalizzare e creare consenso sul concetto di real world object come nuovo protagonista della scena catalografica.

JLIS.it vol. 13, no. 3 (September 2022)
ISSN: 2038-1026 online

DOI: 10.36253/jlis.it-481

Riferimenti bibliografici

Coyle, Karen. 2015. «Coyle's InFormation: Real World Objects». *Coyle's InFormation* (blog). 16 gennaio 2015. http://kcoyle.blogspot.com/2015/01/real-world-objects.html.

Cutter, Charles A. (Charles Ammi). 1876. *Rules for a Dictionary Catalogue*. U.S. Government Printing Office. http://archive.org/details/cu31924029519026.

Gorman, Michael. 2018. *I nostri valori, rivisti: la biblioteconomia in un mondo in trasformazione*. A cura di Mauro Guerrini. Tradotto da Giuliano Genetasio. Firenze: Firenze University Press.

Guerrini, Mauro. 2020. *Dalla catalogazione alla metadatazione: tracce di un percorso*. Collana Percorsi AIB 5. Roma: Associazione italiana biblioteche.

———, a c. di. 2022. «La metadatazione: cos'è?» *Biblioteche oggi* 40 (3): 21–50. https://doi.org/10.3302/0392-8586-202203-021-1.

McCallum, Sally. 2022. «Collocation and Hubs. Fundamental and New Version». *JLIS.It* 13 (1): 45–52. https://doi.org/10.4403/jlis.it-12760.

Riley, Jenn. 2004. «Understanding Metadata». NISO website. 2004. https://www.niso.org/publications/understanding-metadata.

Schreur, Philip E. 2019. «Sinopia: A New Linked-Data Editing Environment Designed for Libraries». In *Metadata and Semantic Research*, a cura di Emmanouel Garoufallou, Francesca Fallucchi, e Ernesto William De Luca, 425–30. Communications in Computer and Information Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-36599-8_39.

Wennerlund, Bodil, e Anna Berggren. 2019. «Leaving Comfort Behind: A National Union Catalogue Transition to Linked Data», 10.

Zeng, Marcia Lei, e Jian Qin. 2016. *Metadata*. 2nd edition. Chicago: Neal-Schuman.

Zhu, Linhong, Majid Ghasemi-Gol, Pedro Szekely, Aram Galstyan, e Craig A. Knoblock. 2016. «Unsupervised Entity Resolution on Multi-Type Graphs». In *The Semantic Web – ISWC 2016*, a cura di Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, e Yolanda Gil, 9981:649–67. Lecture Notes in Computer Science. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46523-4_39.