Proposta per una valutazione automatica della completeness dei metadati nel contesto delle biblioteche digitali

«Dig*Italia*» 2-2020
DOI: 10.36181/digitalia-00023

Matteo Lorenzini - *ETH Zurigo; Fondazione Bruno Kessler; Università degli Studi di Trento*
Marco Rospocher - *Università degli Studi di Verona*
Sara Tonelli - *Fondazione Bruno Kessler*

Grazie all'utilizzo dei metadati è possibile accedere ad un vasto numero di risorse rese disponibili attraverso archivi e biblioteche digitali.
Normalmente i metadati sono strutturati secondo uno schema standardizzato e garantiscono l'interoperabilità e l'identificazione di un oggetto digitale facilitando l'accesso a determinati tipi di risorse. Tuttavia, una bassa qualità dei metadati e delle informazioni che questi rendono disponibili, come ad esempio la mancanza di alcuni elementi rispetto allo schema di metadati utilizzato rappresenta tutt'ora un problema comune a molti repositories.
In questo contributo presentiamo sia il nostro lavoro, allo stato attuale in corso, che ha come scopo la valutazione automatica della qualità dei metadati nelle digital libraries, che un'analisi preliminare sulla metadata completeness delle risorse presenti nell'aggregatore di metadati "CulturaItalia".

1. Introduzione

Negli ultimi anni, nel dominio dei beni culturali, abbiamo assistito ad un considerevole aumento sia delle risorse digitalizzate che, di conseguenza, ad un aumento dell'utilizzo di archivi digitali per la loro fruizione e per il loro mantenimento. Di conseguenza, i metadati sono diventati l'elemento fondamentale attraverso il quale gli utenti possono accedere ad un determinato bacino di informazioni ed aumentare la loro conoscenza riguardo ad un particolare dominio.
Tuttavia, nonostante il massiccio utilizzo dei metadati ed il loro ruolo chiave nel garantire l'accessibilità alle risorse culturali, il processo di controllo e verifica della loro qualità manca ancora di una chiara definizione. In letteratura[1], la *metadata quality* è stata presentata come un modo di misurare quanto un oggetto digitale soddisfi un determinato scopo[2]. In tal senso, il *Metadata Quality Framework* svi-

[1] Pennok 2007.
[2] Secondo Pennok, nel dominio dei beni culturali sono identificabili 3 scopi principali che un oggetto digitale dovrebbe garantire: Presevation, Registering, Discovery.

luppato da Bruce e Hillmann[3] è considerato il cardine sul quale poggia la discussione relativa alla *metadata quality*.

Per valutare la qualità dei metadati il *framework* propone sette metriche qualitative: *Completeness, Accuracy, Conformance, Logical Consistency and Coherence, Accessibility, Timeliness, Provenance*.

Sebbene non vi siano casi d'uso in cui queste metriche siano state applicate nell'analisi di un archivio, sarebbero sicuramente di aiuto nel caso di una valutazione sistematica sui metadati applicandole, per esempio, ad una digital library come Europeana[4].

Facendo riferimento al quadro generale poche sono le casistiche di controllo automatico dei metadati[5]. Possiamo citare il lavoro fatto da Ochoa e Duvall[6] dove, per ciascuna delle sette metriche del *Metadata Quality Framework*, viene ipotizzata una funzione in modo da rendere la valutazione quantitativa e non più qualitativa ed il metodo ipotizzato da Peter Kiraly[7] e parzialmente applicato alla digital library Europeana.

In entrambi i casi non vengono considerati tre importanti fattori:

1. Il processo di creazione dei metadati: spesso svolto manualmente da operatori specializzati seguendo le linee guida messe a disposizione dall'archivio che poi provvederà all'*ingestion*. Per cui, le definizioni dei *metadata elements* usati per la descrizione dell'oggetto potrebbero essere interpretate in maniera diversa a seconda del punto di vista dell'operatore.
2. Processo di aggregazione: il processo di aggregazione di risorse digitali viene fatto quando dei metadati messi a disposizione da uno o più providers sono acquisiti in un unico *repository*. Poiché l'aggregazione dei metadati, nel dominio dei beni culturali, è cresciuta esponenzialmente durante gli ultimi anni, il processo di controllo dei metadati dovrebbe essere ricontestualizzato nell'ottica di verificare informazioni non omogenee tra di loro e su larga scala.
3. Contesto: l'utilizzo dei metadati per descrivere un oggetto varia a seconda del dominio o del contesto in cui viene descritto l'oggetto. Per esempio, un archeologo e un filologo hanno diverse percezioni di un'epigrafe: per il primo, un'epigrafe documenta un reperto archeologico, mentre per il secondo l'epigrafe rappresenta un testo da analizzare.

Anche la definizione di metadato definita da NISO[8] sottolinea l'importanza del contesto descrivendo il metadato come un'entità non statica: hanno diverse interpretazioni e dovrebbero essere considerate in base al loro contesto.

Gli esistenti lavori sulla *metadata quality* presentano inoltre, secondo il nostro pa-

[3] Bruce 2004.
[4] https://www.europeana.eu/en.
[5] Ochoa – Duvall 2009, Ostojic – Sugimoto – Durco 2017.
[6] Ochoa – Duvall 2009.
[7] Kiraly 2015.
[8] https://www.niso.org.

rere, alcune limitazioni per quanto riguarda la granularità dell'analisi in quanto si concentrano o su una specifica metrica oppure focalizzano l'analisi relativamente ad un unico *repository*. Per valutare la *completeness* di un *repository* sono stati presentati quindi due diverse tipologie di approccio:

– La presenza o assenza dei *metadata elements*: affrontata come un problema binario assegnando quindi valore 0 o 1 a seconda della presenza o assenza di un particolare *metadata element*.
– Per ciascun *metadata element* viene attribuito un valore/peso arbitrario a seconda dell'importanza rispetto al profilo di metadati utilizzato.

Questi approcci, anche se consentono al curatore di misurare la qualità del singolo record o, più generalmente, di un intero *dataset*, non considerano se una bassa qualità della risorsa o del *dataset* possa dipendere dalla mancanza di pochi *metadata element* con uno *scoring* alto o dalla mancanza di molti *metadata element* con uno *scoring* basso.
La nostra soluzione che andiamo a proporre si basa sull'assunzione che queste metriche debbano permettere ai curatori di definire in un modo flessibile quali *metadata element* possano essere considerati come più interessanti nella valutazione complessiva della *completeness*, al fine di omogeneizzare questa metodologia di valutazione tra diversi *repositories* e per permettere una più dettagliata analisi dei *metadata elements*.
Questo rappresenta il maggior contributo di questo articolo: proponiamo un modo flessibile per valutare la *completeness* considerando sia i metadati obbligatori che quelli facoltativi di un particolare *metadata schema* cosi come il dominio di una collezione digitale.
Presentiamo inoltre una valutazione della *metadata completeness* prendendo come caso d'uso il *repository* di metadati CulturaItalia[9] evidenziando come, con l'aiuto di una rappresentazione grafica, il nostro approccio possa supportare gli esperti per risolvere il problema della valutazione *metadata quality*.

2. La metodologia proposta

Il nostro obiettivo a lungo termine è quello di sviluppare un *framework* che automaticamente verifichi la qualità dei metadati di un *repository* considerando diverse metriche. Lo sviluppo di questo *framework* è legato allo svolgimento delle seguenti attività:

– Definizione delle metriche di valutazione, catturando lo stato dei *metadati* sia a livello del singolo oggetto digitale (per esempio controllare la qualità dei metadati su un singolo record in un *dataset*) che a livello dell'intero *repository*.

[9] http://www.culturaitalia.it.

– Definizione di algoritmi per calcolare le metriche sopra descritte e, possibilmente, restituire suggerimenti su come riparare la bassa qualità dei metadati.

In questo articolo presentiamo i primi risultati relativi alla *completeness*.
In termini generali, la *completeness* viene calcolata come il rapporto degli elementi usati per descrivere l'oggetto rispetto al profilo di metadati utilizzato. In questa valutazione, diverse variabili dovrebbero essere prese in considerazione: gli elementi che sono obbligatori e quelli che sono opzionali, il contesto ed il dominio della collezione, cosi come le preferenze dei curatori quando viene valutata la *completeness*.

La metodologia che proponiamo viene strutturata come segue:
– Dato un *repository* che deve essere valutato, i *metadata elements* sono divisi in gruppi, rappresentanti la loro importanza (per esempio elementi obbligatori / raccomandati / opzionali)
– Per ciascun oggetto O contenuto nel *repository*, la *completeness score* Cg(O) viene calcolata separatamente per ciascuno dei gruppi (G) definiti: il numero dei *metadata element* utilizzati per descrivere ciascun oggetto, viene diviso per il numero totale dei *metadata element* presenti nel gruppo di cui andiamo a calcolare la *completeness.* Per esempio, se un oggetto viene descritto utilizzando 3 dei 10 elementi obbligatori, il valore della *completeness* sarà 0.3. Lo score della *completeness* sarà dunque un numero reale in una scala da 0 a 1. Più il valore è vicino ad 1, più completa è la descrizione per quello specifico gruppo di metadati.
– La *completeness score* (una per ogni gruppo di metadati) viene calcolata relativamente ad ogni singolo record. Per valutare la completezza generale del *dataset*, per ciascun gruppo di metadati viene definito un grafico separato dove sull'asse X, utilizzando 10 intervalli, viene rappresentato lo score *range* relativo alla *completeness* (0-0.1;0.1-0.2;0.2-0.3;..0.9-1.0) mentre sull'asse Y viene rappresentata la percentuale degli oggetti relativi ad ogni *completeness* score.

La valutazione della *metadata quality* diversificata per gruppi di importanza e di dominio offre al *metadata curator* la possibilità di avere un quadro completo riguardo allo stato generale del dataset o *repository*. Per cui, il curatore, può intervenire sugli oggetti con una bassa qualità dei metadati valutando i diversi fattori che hanno un impatto sul risultato finale.

3. Caso d'uso: CulturaItalia

CulturaItalia è un *repository* di metadati nel dominio dei beni culturali e rappresenta uno degli aggregatori nazionali della digital library europea Europeana.
Gestito dall'ICCU, accoglie circa 4.500.000 risorse che includono immagini, contenuti audio visivi e testuali. I metadati sono resi disponibili attraverso la piattaforma dati.culturaitalia.it[10] utilizzando lo schema di metadati PICO[11], caratterizzato da 94 elementi allineati con Dublin Core. Gli elementi PICO sono stati suddivisi in Obbligatori (8 elementi), raccomandati (10 elementi) e opzionali (76 elementi).
Per questo articolo sono stati presi come casi d'uso due *datasets*: MuseID-Italia (76.828 record) e Regione Marche (90.602 record). Le risorse dei due *datasets* sono per lo più Opere d'Arte Visiva.
Seguendo la metodologia presentata nel paragrafo uno, la suddivisione in gruppi dei metadati è stata fatta prendendo come riferimento le linee guida tecniche[12] ed i documenti di *mapping* resi disponibili attraverso la documentazione tecnica del portale. Sono stati quindi definiti i seguenti 4 gruppi:

– Elementi obbligatori (8 elementi): gli elementi obbligatori del PICO *profile, dc:title, dc:identifier, dc:subject, dc:type, pico:preview, dc:isReferencedBy, dcterms:license, pico:licenseMetadata.*
– Raccomandati (10 elementi): gli elementi raccomandati del PICO *profile* come ad esempio *dc:description, pico:author, dcterms:spatial.*
– Di dominio: gli elementi opzionali del PICO *profile* che sono però rilevanti per uno specifico dominio. Nel caso di Opere d'arte visiva abbiamo identificato i seguenti 11 elementi: *pico:commissioner, pico:matherialAndTechnique, dcterms:created, dcterms:isPartOf, dcterms:alternative, dcterms:modified, dc:contributor, dc:coverage, dcterms:bibliographicCitation, pico:printer, dcterms:replaces.*
– Opzionali: (76 elementi): i rimanenti elementi opzionali del PICO *profile* come per esempio *dcterm:bibliographicCitation, pico:commissioner, pico:performer* ecc.

3.1 Risultati

Prima di analizzare nel dettaglio la *completeness* dei *datasets* in esame, utilizzando la suddivisione in gruppi dei metadati, abbiamo analizzato la frequenza generale di utilizzo degli elementi PICO nelle risorse presenti in MuseID-Italia e Regione Marche. Le figure 1 e 2 rappresentano la percentuale dei *records* nei due *datasets*.

[10] http://dati.culturaitalia.it.
[11] http://www.culturaitalia.it/opencms/export/sites/culturaitalia/attachments/documenti/picoap/picoap1.0.xml.
[12] http://www.culturaitalia.it/opencms/documentazione_tecnica_it.jsp.

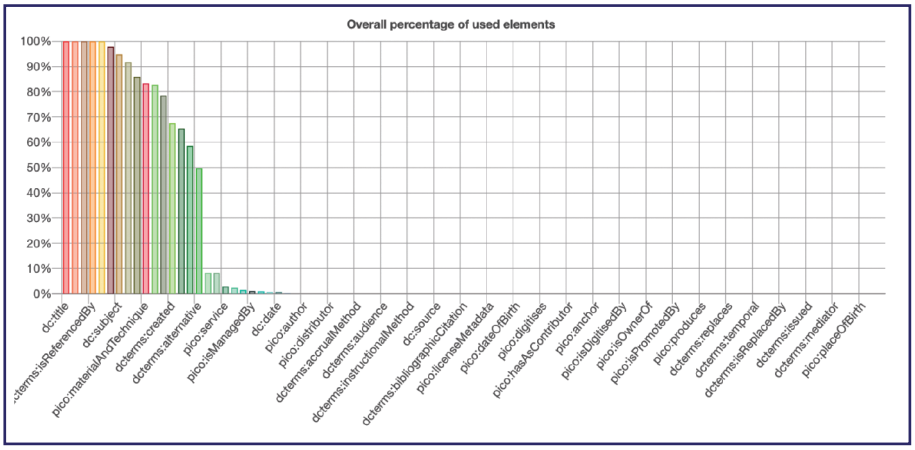

Figura 1. *Percentuale degli elementi PICO per record nel dataset MuseiD-Italia*

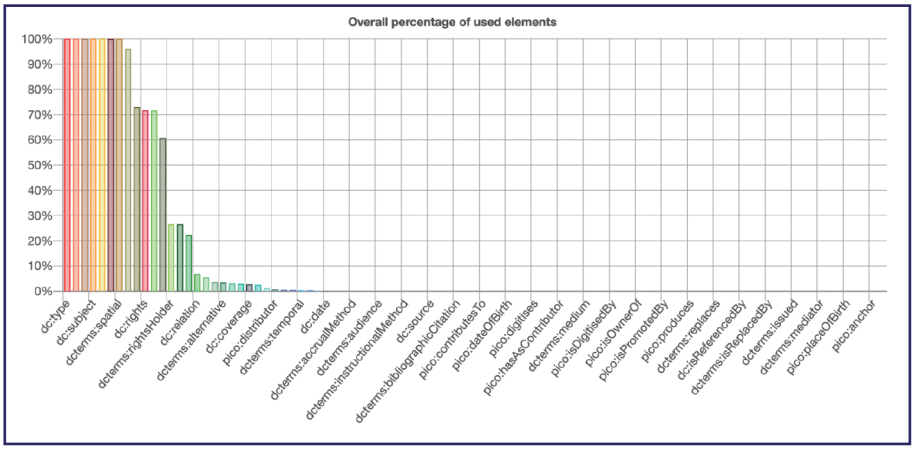

Figura 2. *Percentuale degli elementi PICO per record nel dataset Regione Marche*

Possiamo notare che molti casi (ma non tutti) gli elementi obbligatori sono valorizzati in tutti i record del *dataset*. Diverso è invece il quadro ottenuto dalla verifica degli elementi che caratterizzano i singoli gruppi dove la percentuale degli elementi valorizzati è sostanzialmente più bassa.

Successivamente, utilizzando il raggruppamento definito nella sezione 3, abbiamo analizzato, per ciascun record del *dataset*, 4 diverse *completeness* (una per ciascun gruppo di metadati) e, separatamente, per ciascun gruppo di metadati abbiamo analizzato la distribuzione del grado di *completeness* attraverso un istogramma come descritto nella sezione 2, figure 3 e 4.

I grafici dimostrano che i risultati ottenuti per la *completeness* nei due *datasets* sono generalmente bassi. Per esempio, nel dataset MuseID-Italia (Fig. 3 in alto a sinistra), la maggior parte dei *records* (~92%) ha un valore di *completeness* per i

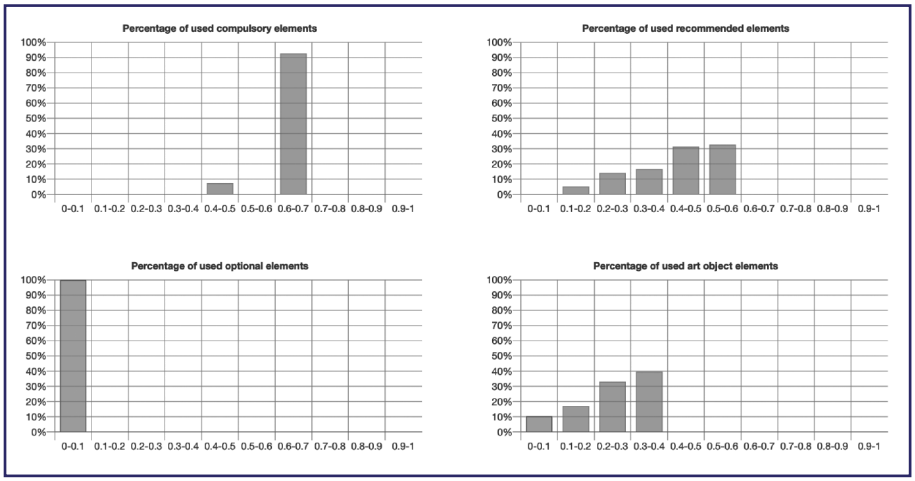

Figura 3. *Completeness per il dataset MuseiD-Italia*

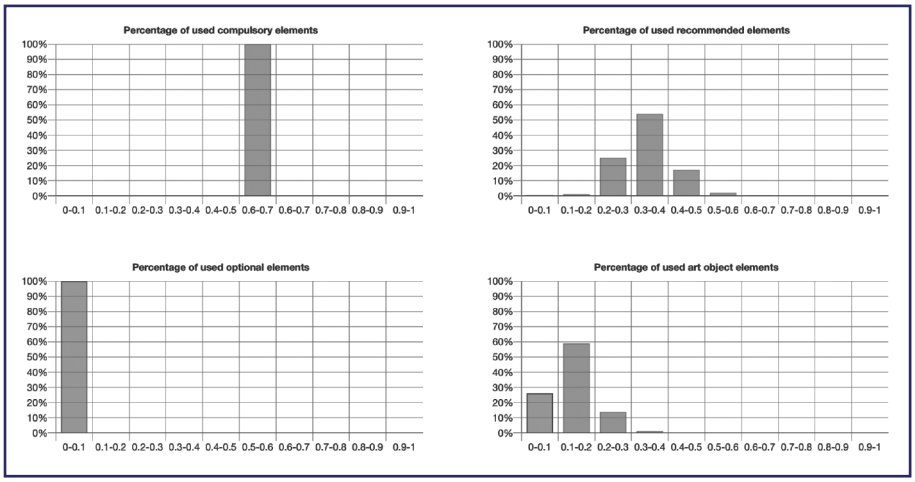

Figura 4. *Completeness per il dataset Regione Marche*

campi obbligatori in un range tra 0.6 e 0.7, nel caso della Regione Marche (Fig. 4 in alto a sinistra) tutti i record raggiungono uno score per gli stessi gruppi di metadati tra 0.5 e 0.6. Lo stesso può essere osservato nella valutazione sul dominio specifico (tra 0.1 e 0.4). I record del *dataset* quindi raramente usano gli elementi opzionali, mentre l'utilizzo degli elementi raccomandati e specifici varia rispetto al dominio.

Abbiamo sottolineato che, grazie ad un'analisi basata su 4 diversi raggruppamenti di metadati, quello che viene restituito al curatore è un'analisi precisa e comprensiva relativa al grado di *completeness* dell'intero *dataset* permettendone anche un confronto rispetto ai 4 raggruppamenti che sono stati definiti.

4. Conclusioni

Applicando concretamente il metodo proposto in due *datasets* come MuseID-Italia e Regione Marche, abbiamo introdotto un metodo innovativo e dettagliato di analizzare i *metadata element* sulla base di diversi raggruppamenti considerando la loro rilevanza rispetto al *dataset*.

In questo modo, il curatore dei metadati può identificare in maniera immediata quali siano i problemi all'interno del *repository* ottimizzando il processo di verifica.

Metadata allow the access to a wide variety of cultural heritage resources made available through repositories, digital libraries and catalogues. Usually taking the form of a structured set of descriptive elements, metadata assist in the identification, location, processing, tracking, preserving, sharing and retrieval of information, while facilitating content and access management. However, low metadata quality, such as the lack of mandatory information, incorrect information or inconsistency, is still an open issue in many repositories. In this paper we present our ongoing work aiming at automatizing the metadata quality analysis, and the preliminary results on metadata completeness for the metadada national aggregator CulturaItalia.

L'ultima consultazione dei siti web è avvenuta nel mese di dicembre 2020

Riferimenti bibliografici

Bruce - Hillmann 2004 — Thomas R. Bruce – Diane I. Hillmann. *The continuum of metadata quality: defining, expressing, exploiting*. In: *Metadata in practice*, ed. by D. I. Hillmann, E. L. Westbrooks. Chicago: American Library association, 2004, p. 238-256

Buonazia – Masci 2007 — Irene Buonazia – Maria Emilia Masci. *Il pico application profile. Un dublin core application profile per il portale della cultura italiana.* Intervento tenuto al Seminario nazionale di studi *Interoperabilità di contenuti e servizi digitali: metadati, standard e linee guida. Roma, 3 aprile 2007*

Buonazia – Masci – Merlitti 2007 — Irene Buonazia – Maria Emilia Masci – Davide Merlitti. *The project of the Italian culture portal and its development. A case study: Designing a Dublin core application profile for interoperability and open distributions of cultural contents*. In: *Openness in Digital Publishing: Awareness, Discovery and Access - Proceedings of the 11th International Conference on Electronic Publishing held in Vienna - ELPUB 2007, Vienna, Austria, June 13-15, 2007*. ELPUB, 2007, p. 393-404

Király 2015 — Péter Király. *A metadata quality assurance framework*. 2015. https://pkiraly.github.io/metadata-quality-project-plan.pdf

Király – Büchler 2018 — Péter Király – Marco Büchler. *Measuring completeness as metadata quality metric in Europeana*. In: IEEE International Conference on Big Data (Big Data), 2018, p. 2711-2720.

Neumaier – Umbrich – Polleres 2016 — Sebastian Neumaier – Jürgen Umbrich – Axel Polleres. *Automated quality assessment of metadata across open data portals*. «Journal of Data and Information Quality (JDIQ)», 1 (2016), p. 1-29.

Ochoa – Duvall 2009 — Xavier Ochoa – Erik Duvall. *Automatic evaluation of metadata quality in digital repositories*. «International journal on digital libraries» 10 (2009), n. 2-3, p. 67-91

Ostojic – Sugimoto – Durco 2017 — Davor Ostojic – Go Sugimoto – Matej Durco. *The curation module and statistical analysis on VLO metadata quality*. Linköping Electronic Conference Proceedings, 2017, p. 90-101. <https://ep.liu.se/ecp/136/007/ecp17136007.pdf>

Pennok 2007 — Maureen Pennok. *Digital curation: a life-cycle approach to managing and preserving usable digital information*. «Library & Archives», (2007), p. 34-45.

Tani – Candela – Castelli 2013 — Alice Tani – Leonardo Candela – Donatella Castelli. *Dealing with metadata quality: The legacy of digital library efforts.* «Information Processing & Management», 6 (2013), p. 1194-1205.

Valentine et al. 2017 — Charles Valentine – Juliane Stiller – Péter Király – Werner Bailer – Nuno Freire. *Data Quality Assessment in Europeana: Metrics for Multilinguality.* In TDDL/MDQual/Futurity, 2017. <http://ceur-ws.org/Vol-2038/paper6.pdf>.