

Valutazione qualità

Gabriele Pisciotta

5/12/2020

In questa sperimentazione vogliamo valutare la qualità dei match ottenuti attraverso le query per bibref, sia verso l'indice locale sia verso Crossref.

Configurazione della sperimentazione

- Selezionato 5 documenti citanti e relative citazioni (output di BEE)
- Creato un CSV estraendo ogni coppia (DOI, bibref) presente nei 5 documenti citanti
- Per ogni riga del CSV, prendo la bibref ed effettuo la query sia verso l'indice locale, sia verso Crossref, salvando i risultati in altre due colonne del CSV.
- Annoto manualmente il DOI nei casi:
 - DOI esplicito mancante
 - DOI discordanti tra locale e remoto
 - DOI discordanti tra locale ed esplicito
 - DOI discordanti tra remoto ed esplicito.
- Se non è stato possibile trovare un DOI manualmente, inserisco un “???”.
- Escludo dal file CSV le righe appartenenti ai seguenti casi:
 - il doi manuale non è stato possibile trovarlo (valore: ???)
 - il doi trovato manualmente non è incluso nel dump di Crossref (e quindi non sarebbe possibile trovarlo!)

Elementi valutati

Saranno valutati i seguenti valori al fine di caratterizzare la qualità dei match:

- numero di volte in cui il DOI locale è corretto e quello remoto no
- numero di volte in cui il DOI remoto è corretto e quello locale no
- numero di volte in cui sono entrambi errati e identici
- numero di volte in cui sono entrambi errati ma differenti
- numero di volte in cui sono entrambi corretti e identici
- percentuale di corrispondenza tra DOI locale e DOI manuale (quanta correttezza usando l'indice)
- percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta correttezza usando le API remote)

- percentuale di corrispondenza tra DOI remoto e DOI locale (quanta corrispondenza tra l'indice e le API)
- percentuale entrambi corretti
- percentuale entrambi errati e identici
- percentuale entrambi errati ma differenti

Riguardo il caso in cui sono entrambi errati siamo infine interessati a:

- percentuale entrambi errati e identici (su tutti gli errati)
- percentuale entrambi errati ma differenti (su tutti gli errati)

Risultati

I seguenti risultati sono stati valutati su un numero di 203 record a seguito di un preprocessing volto ad escludere le righe appartenenti a casistiche indesiderate, come definito nella configurazione della sperimentazione.

- numero di volte in cui il DOI locale è corretto e quello remoto no: 7
- numero di volte in cui il DOI remoto è corretto e quello locale no: 2
- numero di volte in cui sono entrambi errati e identici: 3
- numero di volte in cui sono entrambi errati ma differenti: 1
- numero di volte in cui sono entrambi corretti e identici: 189
- percentuale di corrispondenza tra DOI locale e DOI manuale (quanta - correttezza usando l'indice): 97.03%
- percentuale di corrispondenza tra DOI remoto e DOI manuale (quanta - correttezza usando le API remote): 94.55%
- percentuale di corrispondenza tra DOI remoto e DOI locale (quanta - corrispondenza tra l'indice e le API): 95.05%
- Percentuale entrambi corretti: 93.56%
- Percentuale entrambi errati e identici: 1.48%
- Percentuale entrambi errati ma differenti: 0.49%

Circa gli errati:

- percentuale entrambi errati e identici (su tutti gli errati): 75.0%
- percentuale entrambi errati ma differenti (su tutti gli errati): 25.0%