FONDAZIONE BRUNO KESSLER

What NLP can do for Metadata Quality

The Case of Descriptions in Cultural Heritage Records

Sara Tonelli –

joint work with Matteo Lorenzini and Marco Rospocher

Fondazione Bruno Kessler, Trento

satonelli@fbk.eu

Metadata Quality

Recognised as important in the literature, yet there is no agreement on what metadata quality is

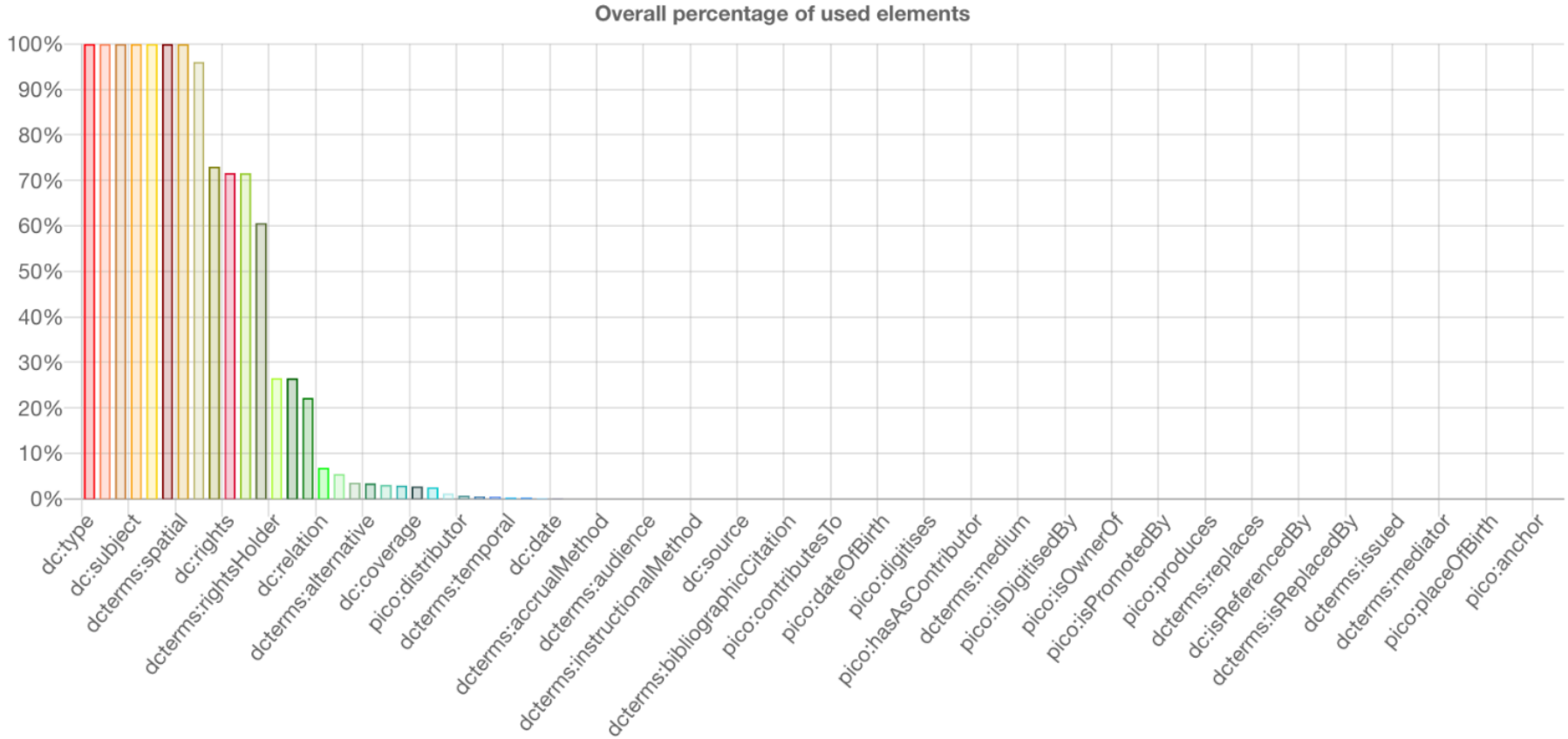
Intuitively defined as "fitness for use", its understanding changes from one community to the other

Metadata quality has been decomposed into several dimensions, which should all be taken into account to consider metadata of good quality in a quality assessment framework

Frameworks for Quality Assessment

Framework	Parameters	Metrics
Bruce and Hillman (2004)	7	n.a.
Ochoa and Duval (2009)	7	13
Stvilia et al. (2009)	22	41
Hughes (2004)	7	7
Bethard et al., (2009)	7	7
Candela, Athanasopoulos et al. (2011)	20	0
Moreira et al. (2009)	10	10

Example: Completeness

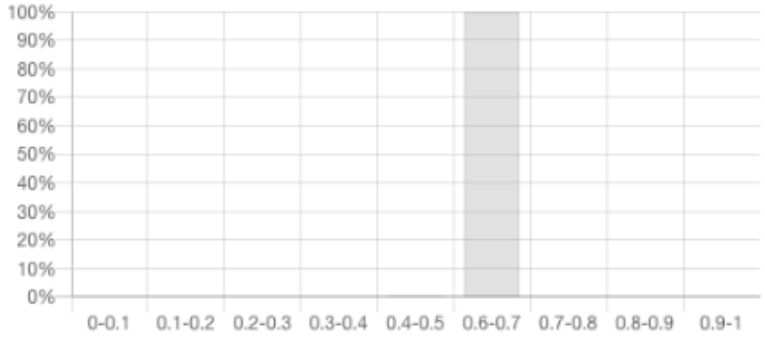


Percentage of records in *Regione Marche* dataset
for each metadata field

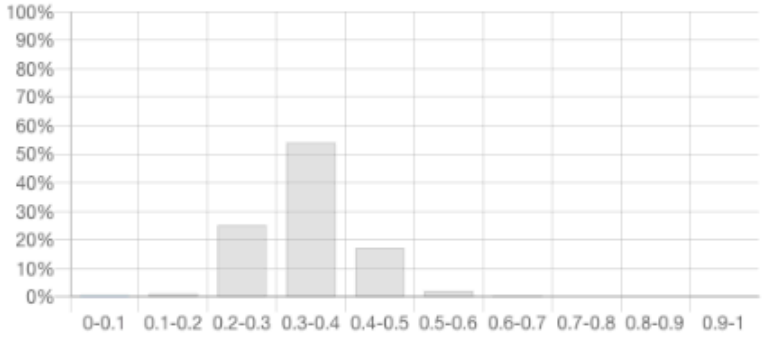
FONDAZIONE
BRUNO KESSLER

Example: Completeness

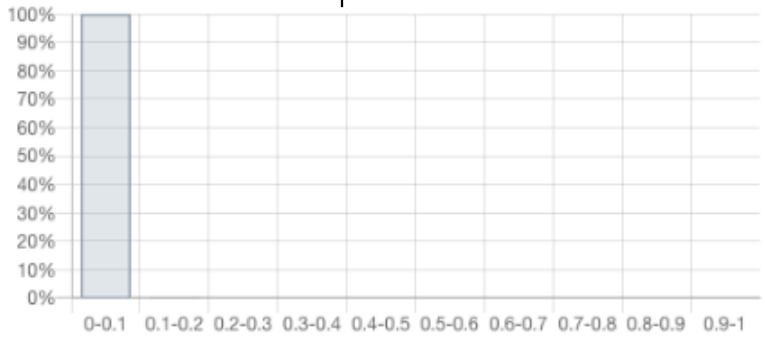
Mandatory



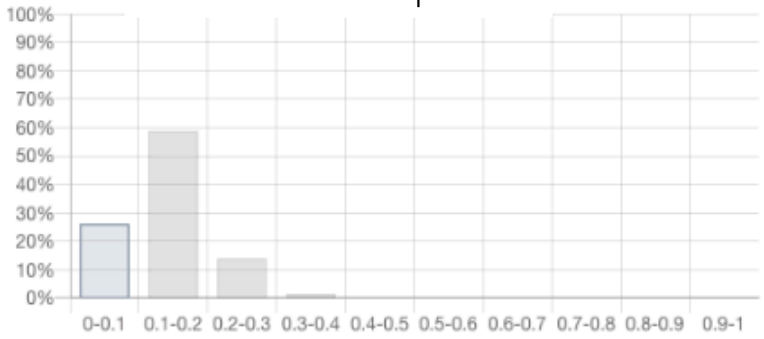
Recommended



Optional



Domain-specific



FONDAZIONE BRUNO KESSLER

Metadata in *Regione Marche* dataset divided into categories

Our Contribution: Accuracy

"Metadata should be accurate in the way it describes objects. The information provided in the value needs to be correct and factual" (Bruce and Hillman, 2004)

Focus on *dc:description*

Goal: Automatically assess the low or high-quality of information in the *dc:description*

FONDAZIONE
BRUNO KESSLER

Guidelines for accurate descriptions

Guidelines provided by Istituto Centrale per il Catalogo e la Documentazione (ICCD):

"The object typology and shape must be described. To describe the object, the cataloguer must refer to the vocabularies provided by ICCD. The description of the subject must report the iconographic and decorative settings. For example, the characters of the depicted scene in a painting and their attribution"

FONDAZIONE
BRUNO KESSLER

Our Contribution: Accuracy

"Dipinto entro cornice lignea verniciata ocra con bordo interno dorato. Amedeo III è raffigurato di profilo in armatura scura con ceselli in oro, mascheroni dorati sulle spalle e sull'elmo, cimiero con piume rosse e bianche. Nella parte inferiore del dipinto fascia con iscrizione a caratteri stampatello.. Personaggi: Amedeo III di Savoia"

FONDAZIONE
BRUNO KESSLER

Our Contribution: Accuracy

"Congdon si è raramente dedicato al disegno come forma espressiva autonoma, così la mole di disegni raccolti sui taccuini non sono altro che appunti visivi presi durante numerosi viaggi. In questo senso non è possibile, se non raramente, assegnare al singolo disegno un'opera finita direttamente corrispondente, così questi disegni non vengono nemmeno ad essere schizzi preparatori. La sommatoria di tutti i disegni relativi a un luogo danno origine a una serie di dipinti che non hanno un corrispettivo oggettivo nei disegni stessi. Tutto questo giustifica la presenza degli appunti all'interno delle immagini (colori, sfumature e spiegazioni di vario genere). Nel caso probabile veduta di Napoli eseguita durante un viaggio del 1951"

Research Questions

RQ1: Is it possible to effectively use NLP and machine learning to assess the quality of cultural heritage descriptions?

RQ2: What is the impact of the domain using automatic quality control?

RQ3: How many annotated instances are needed to create enough training data to automatically assess description quality?

FONDAZIONE
BRUNO KESSLER

Dataset Creation

Italian digital library "Cultura Italia", the Italian aggregator of the European Digital Library Europeana, around 4 million records

Using the *dc:description* element from Dublin Core, 110,000 descriptions have been collected, belonging to *Visual Art Works*, *Archaeology* and *Architecture*

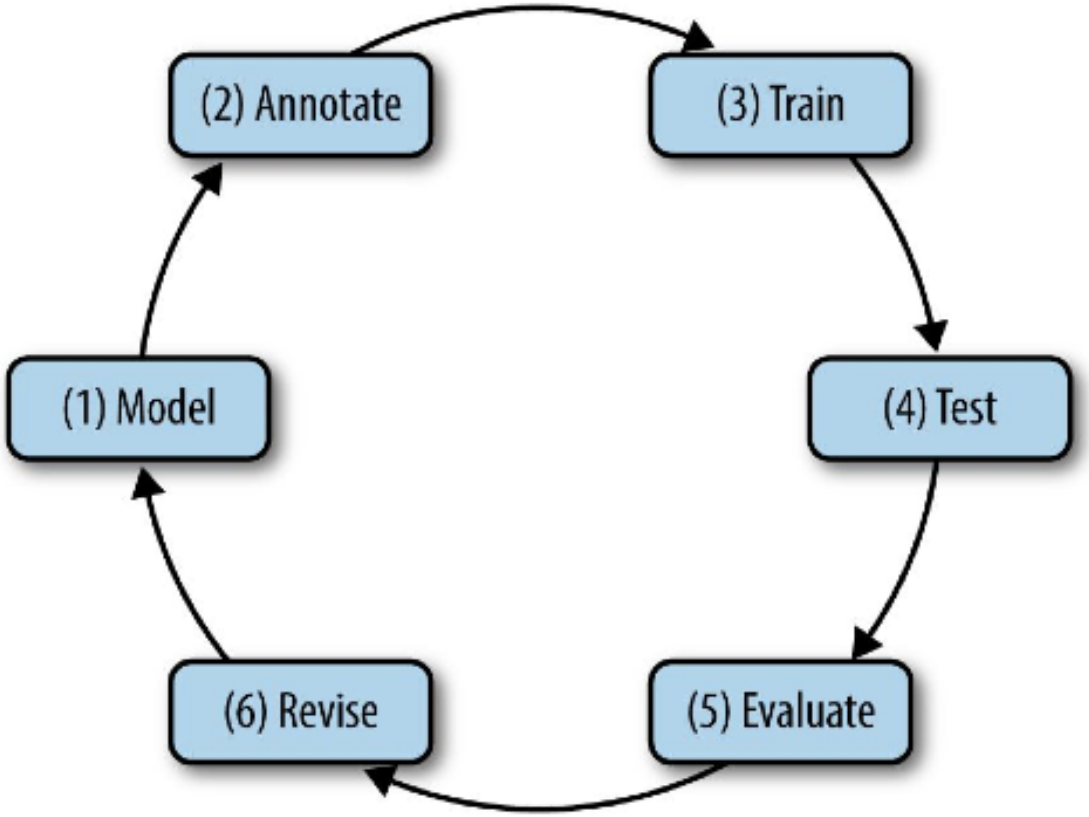
The descriptions have been labeled as "high-quality" or "low-quality" by a domain expert

Dataset Statistics

Dataset	High-Quality	Low-quality	Total
Visual Art Works	30.390	29.611	60.001
Archaeology	19.447	10.803	30.250
Architecture	12.761	7.023	19.784
Overall dataset	62.598	47.437	110.035

FONDAZIONE BRUNO KESSLER

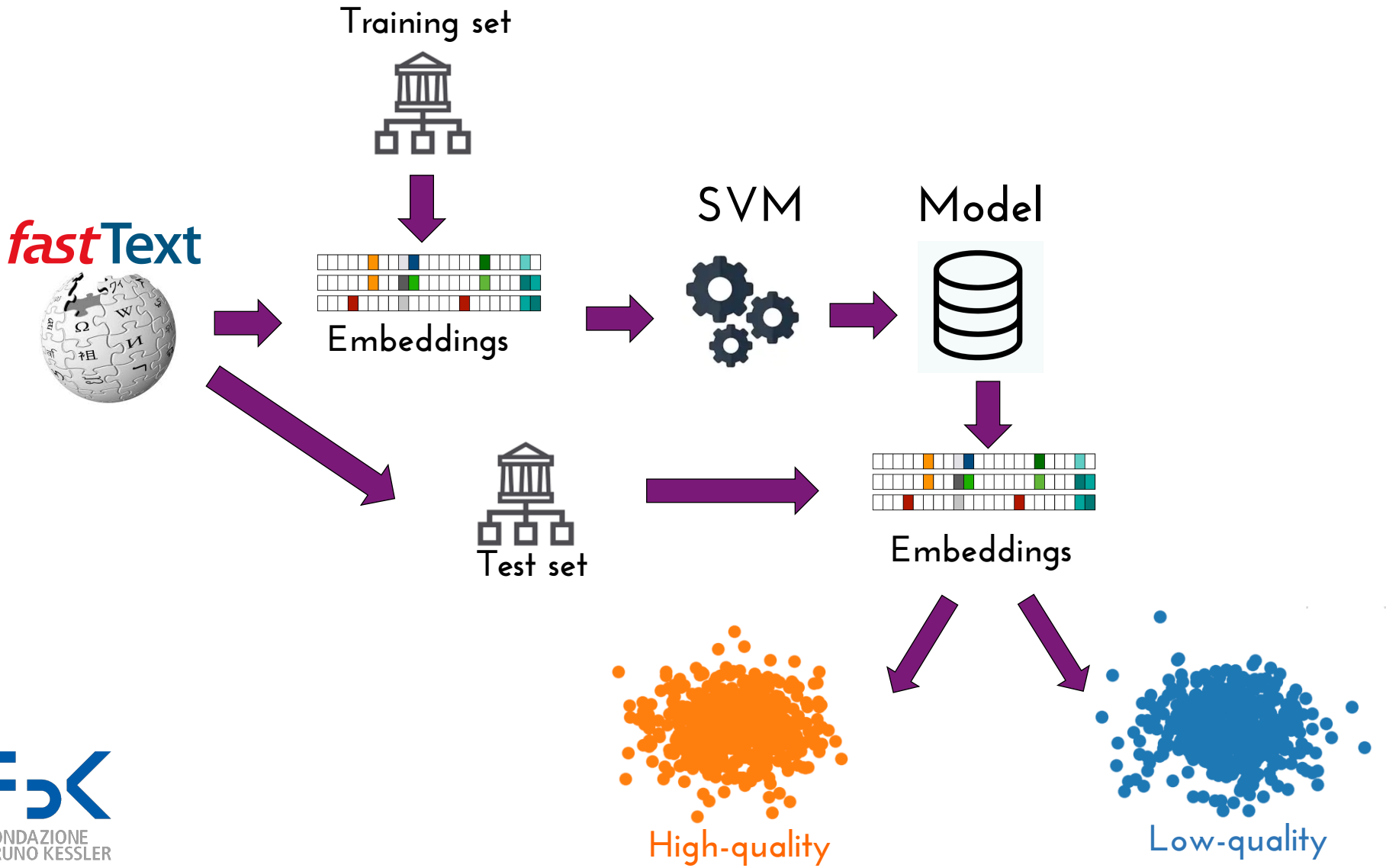
Supervised classification



MATTER cycle: from model to annotated data to automatic systems

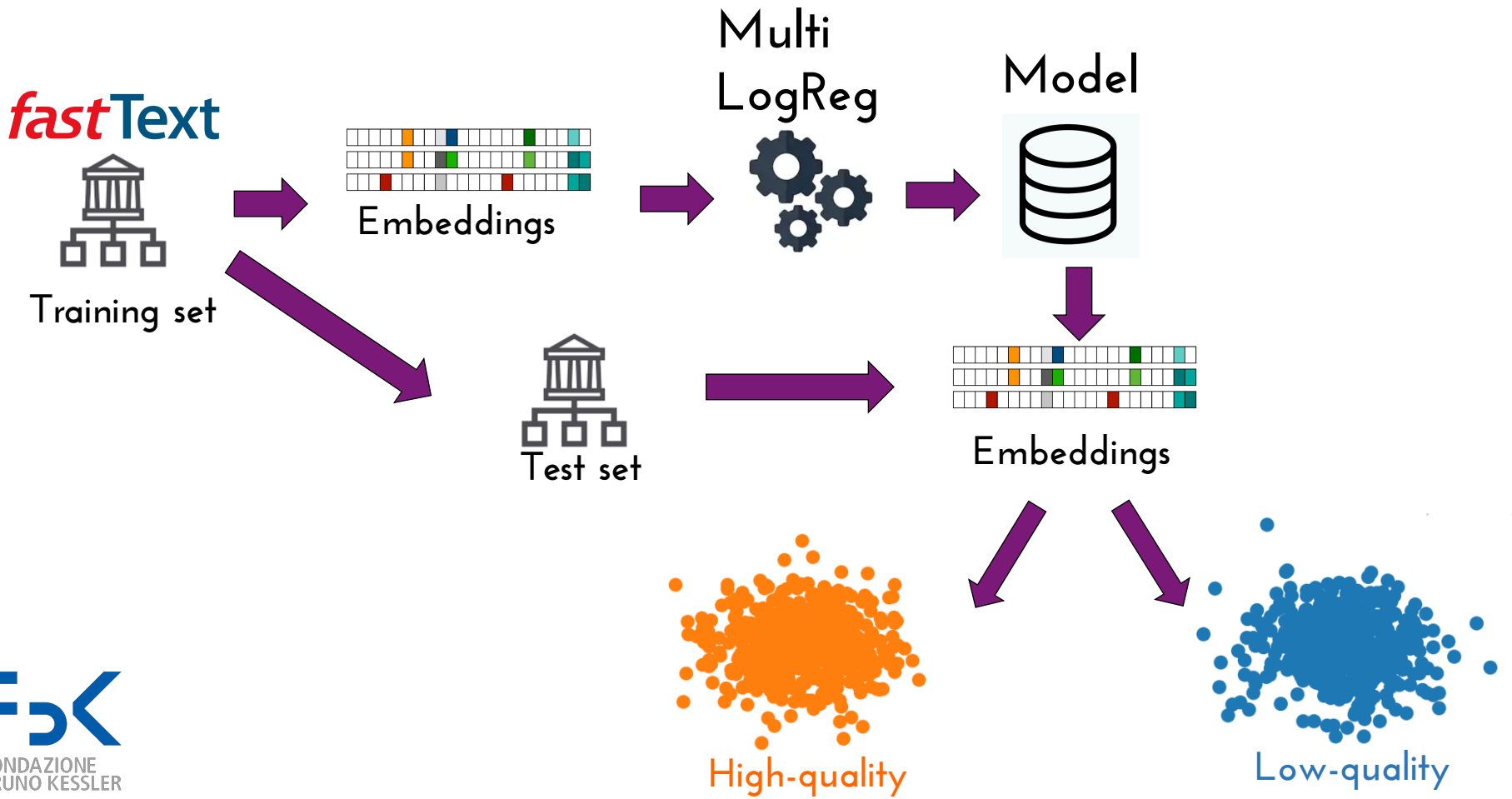
(Pustejovsky and Stubbs, 2012)

Classification Framework (1)



FONDAZIONE BRUNO KESSLER

Classification Framework (2)



Results for In-Domain Classification

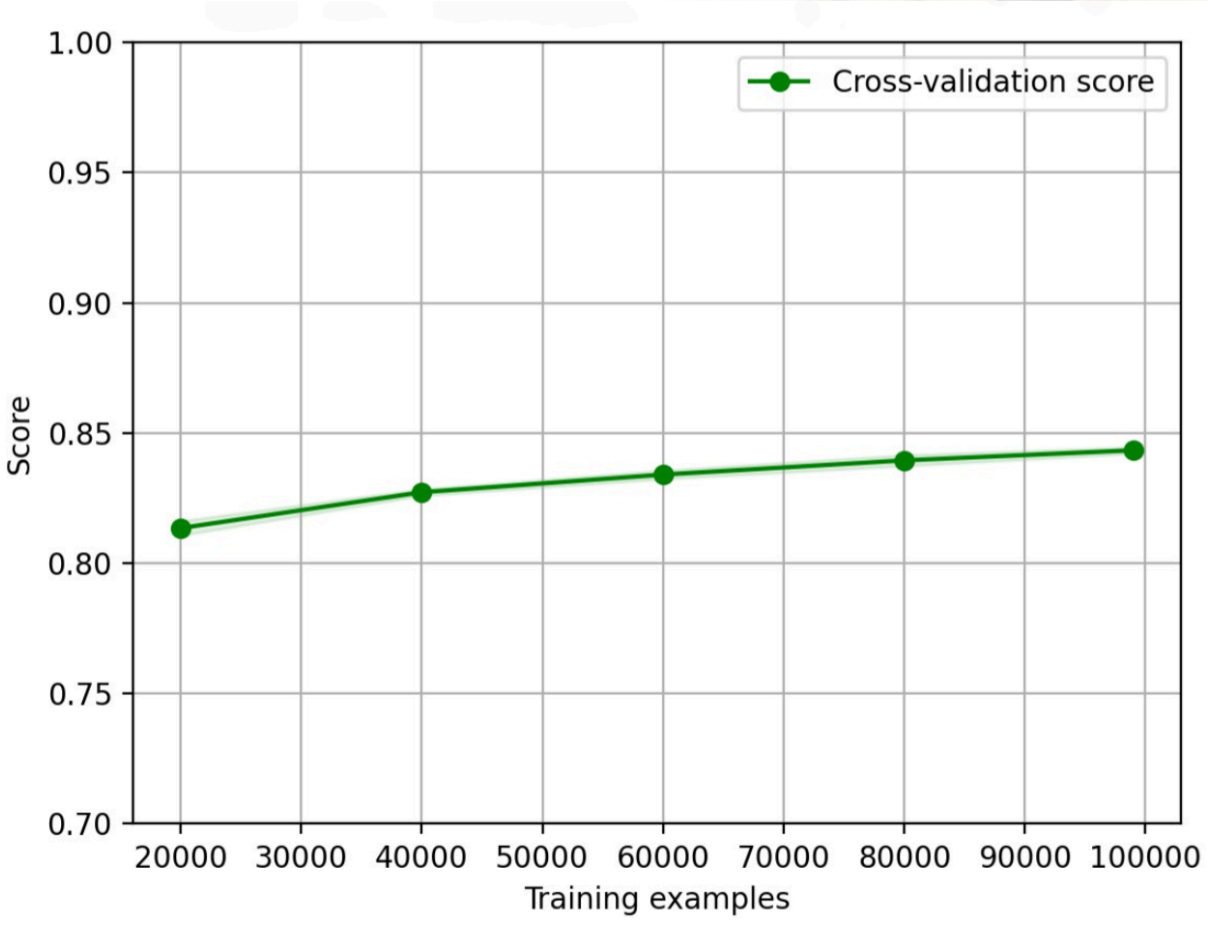
Domain	System	F1	Accuracy
Visual Art Works	SVM	0.86	0.86
	MLR	**0.93**	**0.93**
	Baseline (descr. length)	0.49	0.49
Archaeology	SVM	0.86	0.87
	MLR	**0.93**	**0.93**
	Baseline	0.52	0.65
Architecture	SVM	0.86	0.87
	MLR	**0.93**	**0.93**
	Baseline	0.70	0.75

Results for Cross-Domain Classification

Train	Test	F1	Accuracy
Archaeology + Architecture	Visual Art Works	0.53	0.59
Visual Art Works + Architecture	Archaeology	0.69	0.69
Visual Art Works + Archaeology	Architecture	0.62	0.71

FONDAZIONE BRUNO KESSLER

How Many Training Instances?



FONDAZIONE
BRUNO KESSLER

Lessons Learnt

Classification with in-domain data yields much better results than cross-domain

Domain expertise to create training data is necessary

Explicitly modelling external knowledge is not needed for the classification task (embeddings are enough), although it does not provide insights into the characteristics of good and bad descriptions

Descriptions with Latin and Greek terms are usually misclassified

Conclusions

RQ1: Is it possible to effectively use NLP and machine learning to assess the quality of cultural heritage descriptions? ✔

RQ2: What is the impact of the domain using automatic quality control? ✔

RQ3: How many annotated instances are needed to create enough training data to automatically assess description quality? ✔

FONDAZIONE
BRUNO KESSLER

Thanks for your attention!

Sara Tonelli, satonelli@fbk.eu

Digital Humanities Group
Fondazione Bruno Kessler,
Trento