ECLAP2013

Metadata Quality assessment tool for Open Access Cultural Heritage institutional repositories

Emanuele Bellini, Paolo Nesi

Problem

a) Low metadata quality affects the discover, find, identify, select, obtain - ability of digital resources.

b) Identifying and fixing metadata quality issues is a really time consuming task. With a significant number of records this task might be impossible to execute.

Factors

The creation of metadata automatically or by authors who are not familiar with commonly accepted cataloguing rules, indexing, or vocabulary control can create quality problems. Mandatory elements may be missed or used incorrectly. Metadata content terminology may be inconsistent, making difficult to locate relevant information.

Objective of the work

Definition of

a) A community driven **Metadata Quality Profile** and related quality dimensions able to be assessed through automatic processes

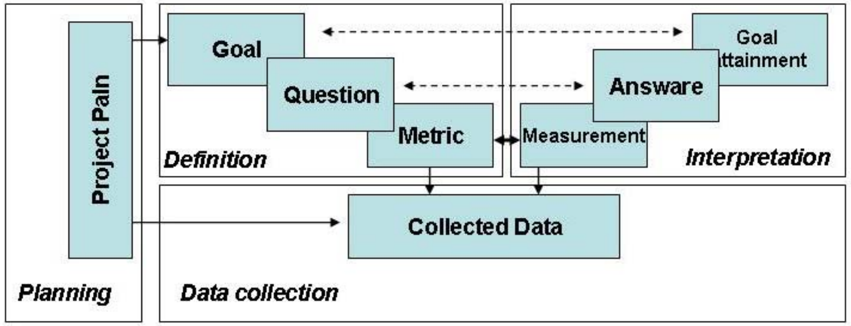
b) A set of **High Level and Low level metrics** to be used as statistical tool for assessing and monitoring the IR implementation in terms of metadata quality, trustworthiness and standard compliance

c) A set of **measurement tools** to asses the defined metrics

d) A Metadata Quality assessment service prototype for automatic evaluation and report

Methodology

Goal Question Metric (GQM) Approach



1) **Planning phase** : Metadata Quality profile definition
2) **Definition phase**: High Level Metrics (HLM) and Low Level Metrics (LLM) definition
3) **Data collection phase**: Measurement Plan definition (ISO/IEC 15939 Measurement Information Model)
4) **Interpretation phase**: Look at the measurement results in a post-mortem fashion. According to the ISO/IEC 15939 this phase foresees the check against thresholds and targets values to define the quality index of the repository

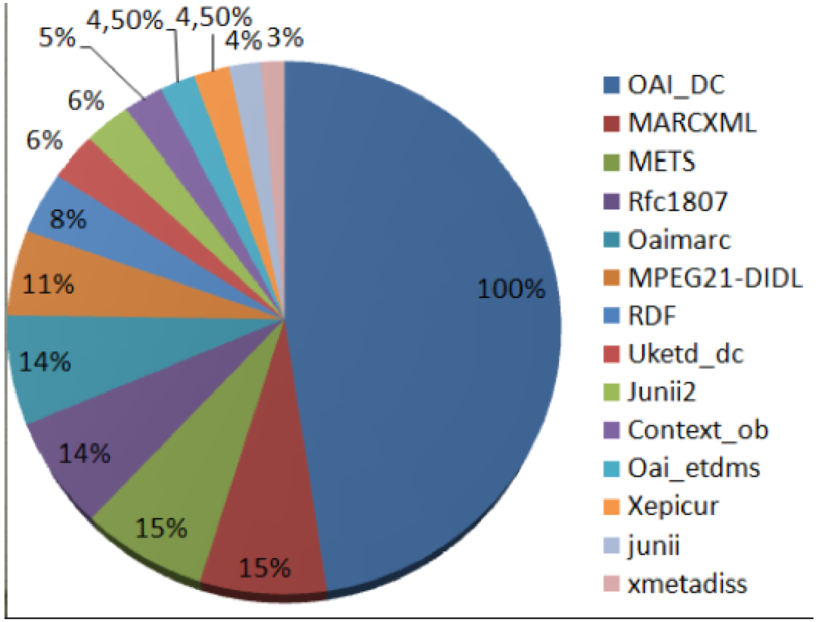
Prototype implementation

Open Archive metadata quality issues analysis

Research conducted over 1200 OA-IR – more than 15M of records analysed

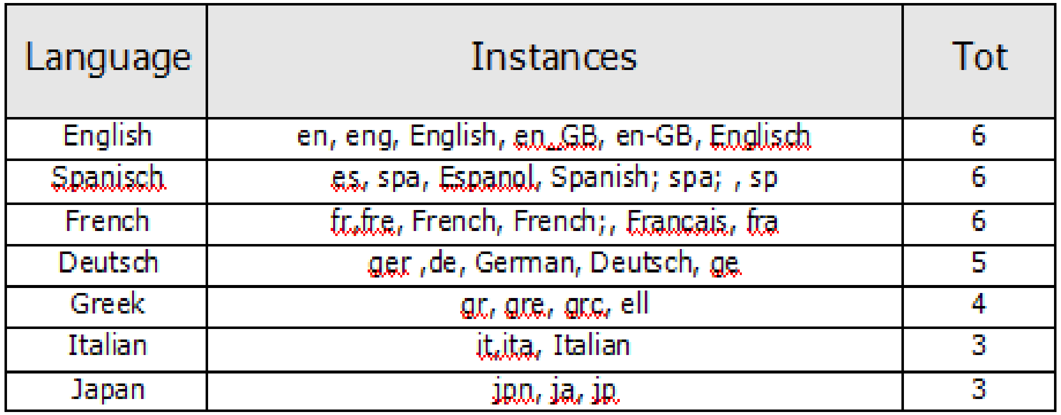
Fragmentary Landscape

More than 100 different metadata sets



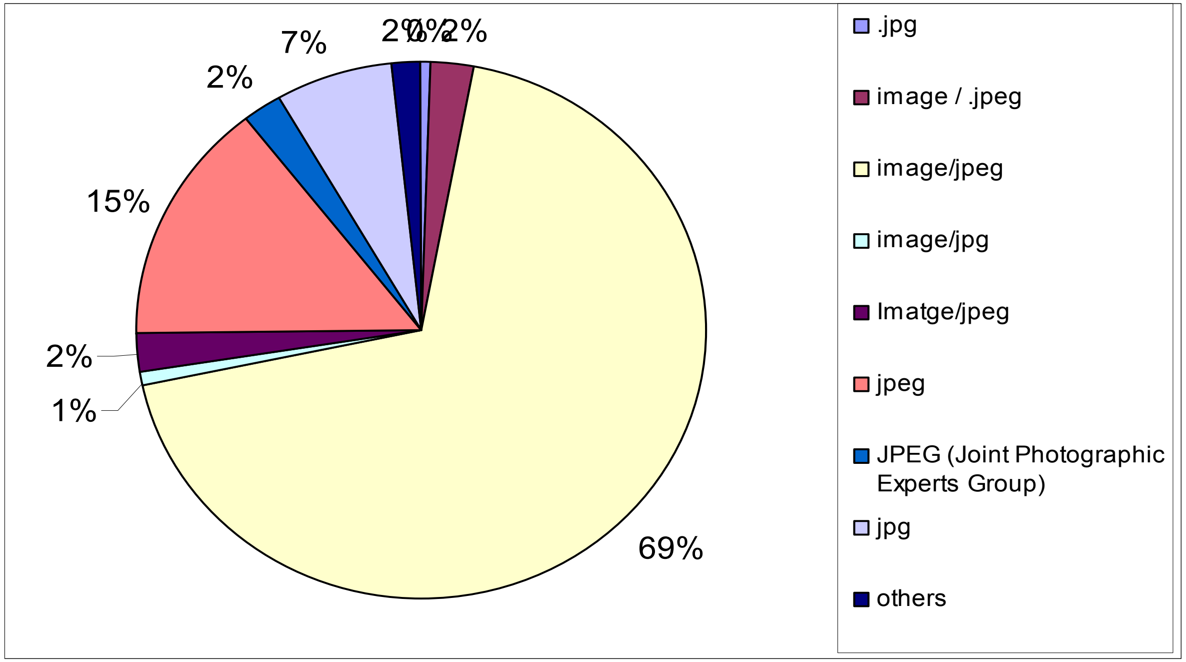
Open Archive metadata quality issues analysis

Language field



Language	Instances	Tot
English	en, eng, English, en_GB, en-GB, Englisch	6
Spanisch	es, spa, Espanol, Spanish; spa; , sp	6
French	fr,fre, French, French;, Francais, fra	6
Deutsch	ger ,de, German, Deutsch, ge	5
Greek	gr, gre, grc, ell	4
Italian	it,ita, Italian	3
Japan	jpn, ja, jp	3

Type field



<u>Wrong values</u>

Collected MIME types contains wrong coding in more than 10% of cases

Open Archive metadata quality issues analysis

- Lack of awareness of recommended open standards

- Difficulties in implementing standards in some cases, due to lack of expertise, immaturity of the standards, or poor support for the standards

- Software tools and interfaces not suitable

- Not well defined duties (which department will be in charge to the IR), publication workflow, rules, policies and responsibilities in the institutions that aims to set up an IR

- Lack of fund and/or human resources for managing IRs

Metadata quality requirements

Metadata quality -> “fitness for use” in a particular typified task/context

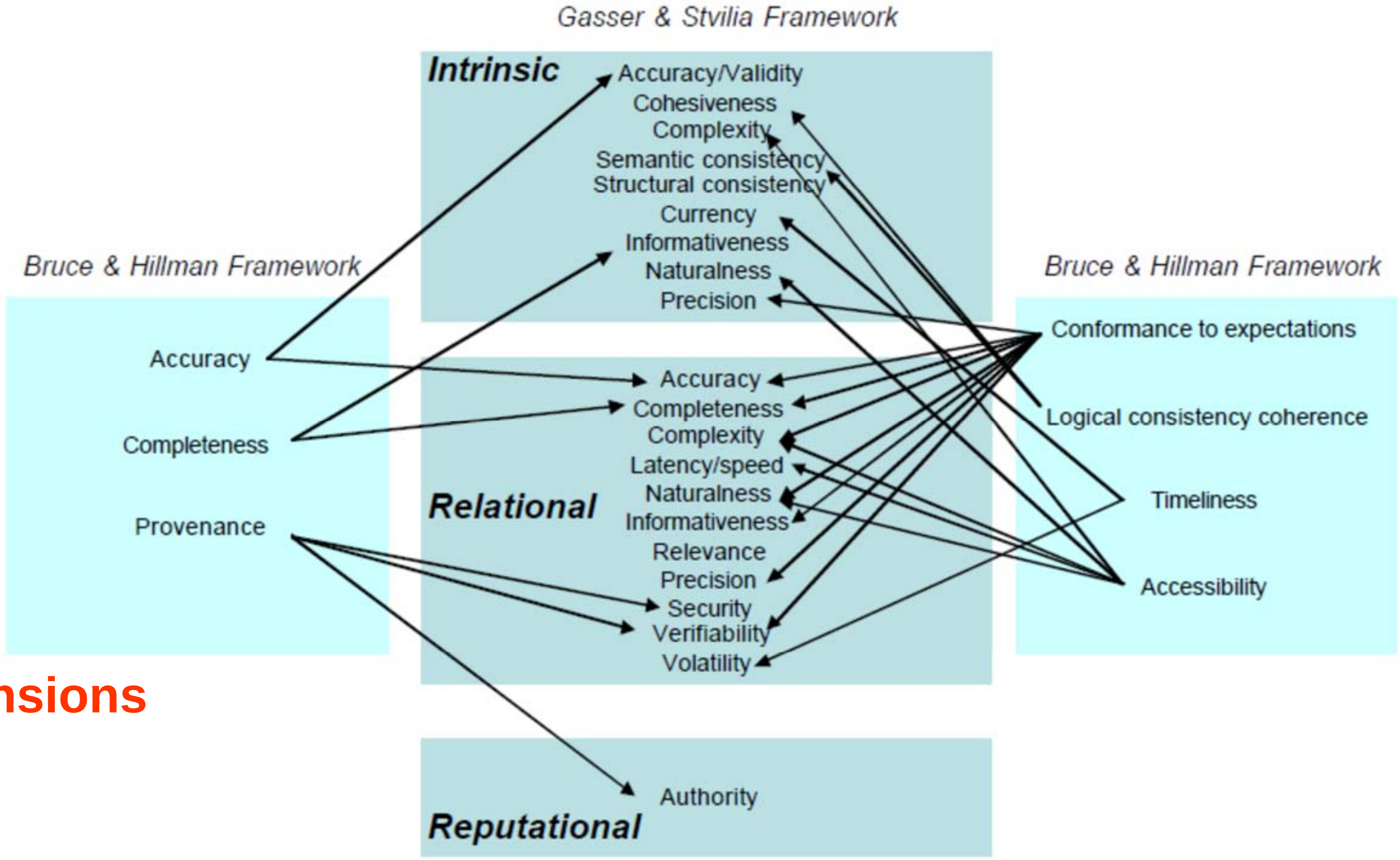
IFLA FRBR model.

Using the descriptive metadata:

- to find materials that correspond to the user’s stated search or discovery criteria
- to identify a resource and to check that the document described in a record corresponds to the document sought by the user, or to distinguish between two resources that have the same title
- to select a resource that is appropriate to the user’s needs (e.g., to select a text in a different language or version)
- to obtain access to the resource described (e.g. to access in a reliable way to an online electronic document stored on a remote computer)

Metadata Quality Frameworks

- NISO *Framework of Guidance for Building Good Digital Collections*
- ISO/IEC 9126
- ISO25000 SQuaRE
- ISO/IEC 14598
- Bruce & Hillman
- Stvilia et al.
- Moen et al



Gasser & Stvilia Framework

Bruce & Hillman Framework

Accuracy

Completeness

Provenance

Intrinsic

Accuracy/Validity

Cohesiveness

Complexity

Semantic consistency

Structural consistency

Currency

Informativeness

Naturalness

Precision

Relational

Accuracy

Completeness

Complexity

Latency/speed

Naturalness

Informativeness

Relevance

Precision

Security

Verifiability

Volatility

Reputational

Authority

Bruce & Hillman Framework

Conformance to expectations

Logical consistency coherence

Timeliness

Accessibility

Actually many dimensions are not computable automatically

OA community driven Quality Profile

Data Filtering

Questionnaire results

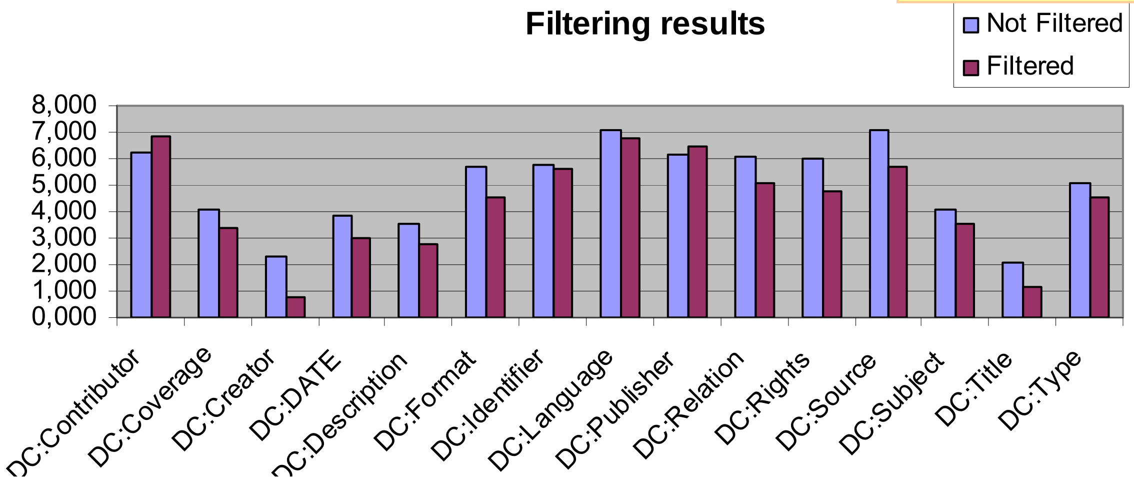
Researchers 20,6%,
Professors 12,7%,
ICT experts15,9%,
Archivists15,9%,
Librarians 25,4%
Students 9,5%

Critical target: Students can represent a “noise”.

Low level of knowledge: The 17% of the responders stated their knowledge of the DC schema is less then 5 (1 to 10 Never worked with metadata

•The work 6,3% of the responders does not include the definition and use of metadata

Never dealt with metadata quality: The 11,1% of the responders has never dealt with the quality of metadata



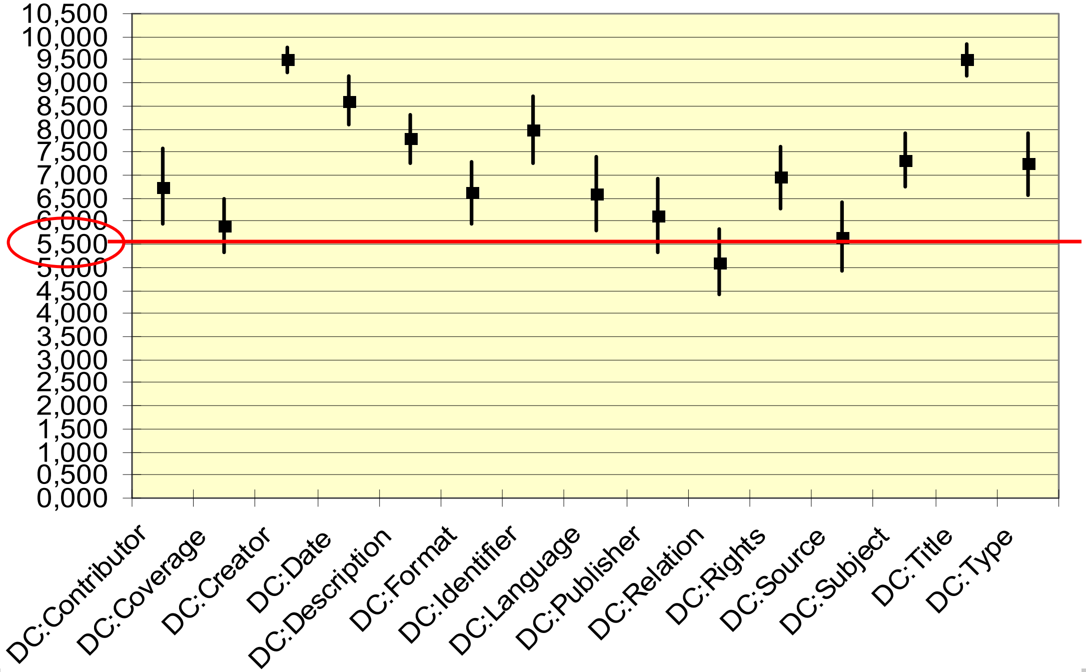
Quality profile

Filed importance from: 1 (the field can be omitted without affect the use of the record) to 10 (absolutely mandatory, the lack of the field makes the record totally unusable).

Range from 1 to 5,5 is considered as not important,

a) The quality assessment on the field *f* can be avoided if the Avg weight **is 5,5 or less**

b) The quality assessment on the field *f* can be avoided if the difference between the AVG weights and the level of confidence is 5,5 or less.



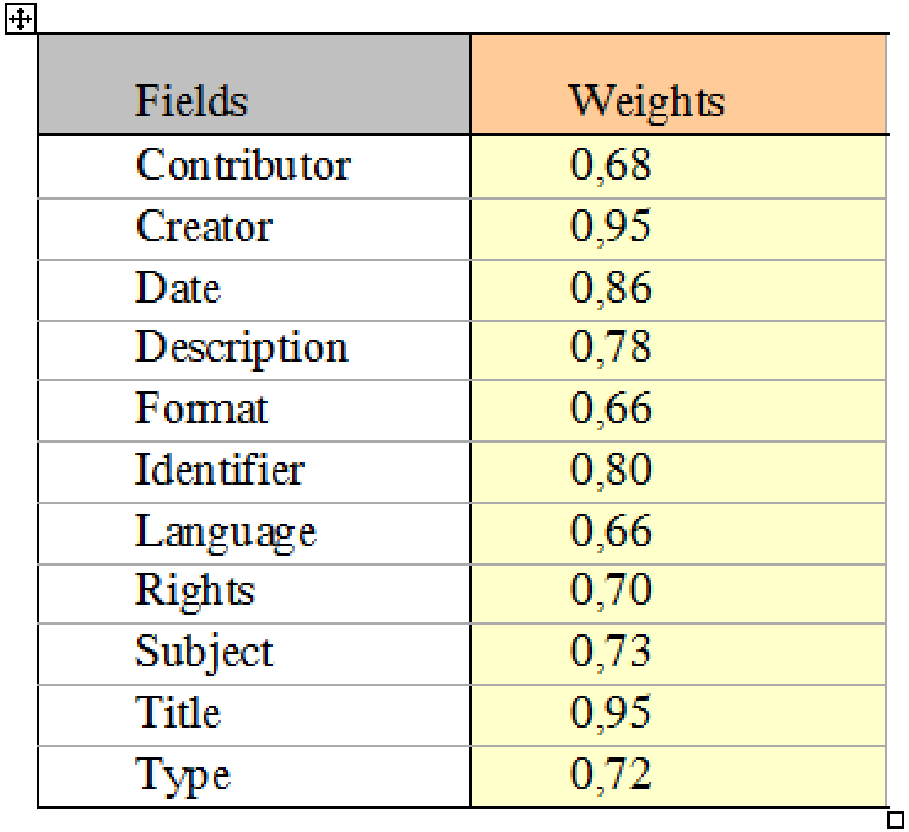
Field selection results

Coverage
Publisher
Relation
Source

Quality profiles

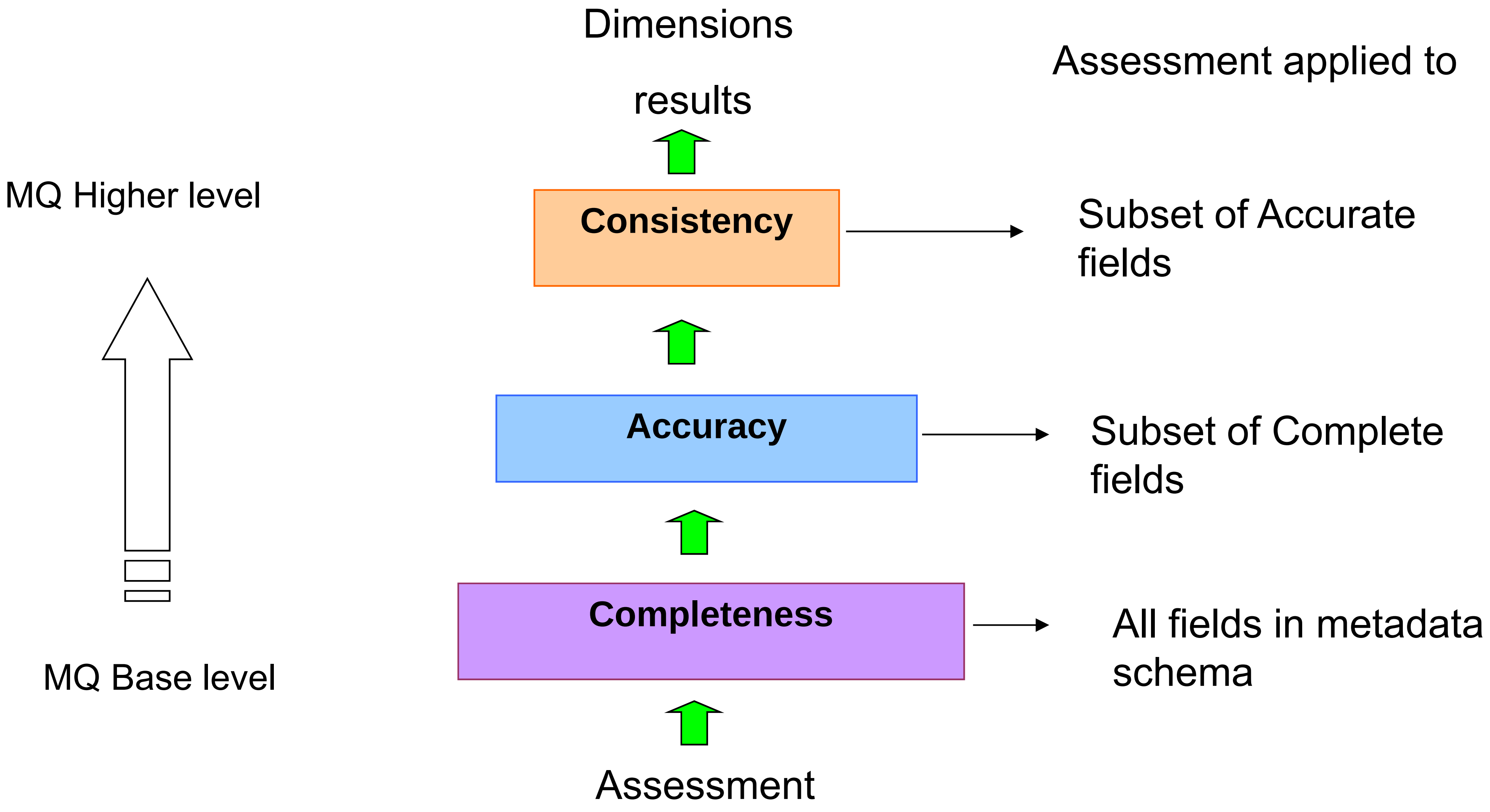
Each field has a different level of relevance in a record

The relevance weights assigned to each field are the normalized Averages of the weights assigned by the AO experts



Fields	Weights
Contributor	0,68
Creator	0,95
Date	0,86
Description	0,78
Format	0,66
Identifier	0,80
Language	0,66
Rights	0,70
Subject	0,73
Title	0,95
Type	0,72

High Level Metrics (HLM)



Dimensions

results

Assessment applied to

MQ Higher level

Consistency

Subset of Accurate fields

Accuracy

Subset of Complete fields

Completeness

All fields in metadata schema

MQ Base level

Assessment

High Level Metrics (HLM) examples

Completeness: If a filed is empty or not.

Accuracy:- there are not typographical errors in the free text fields,
- the values in the fields are in the format defined by standard of reference.
(e.g. ISO639-1 standard for the DC:language)

Consistency: no logical errors
(e.g. a resource results “published” before to be “created”, MIME type declared is different respect to the real bitstream associated, the language of the document if different to the language expressed in the metadata field DC:language the link to the digital objects is broken, etc).

Low Level Metrics (LLM)

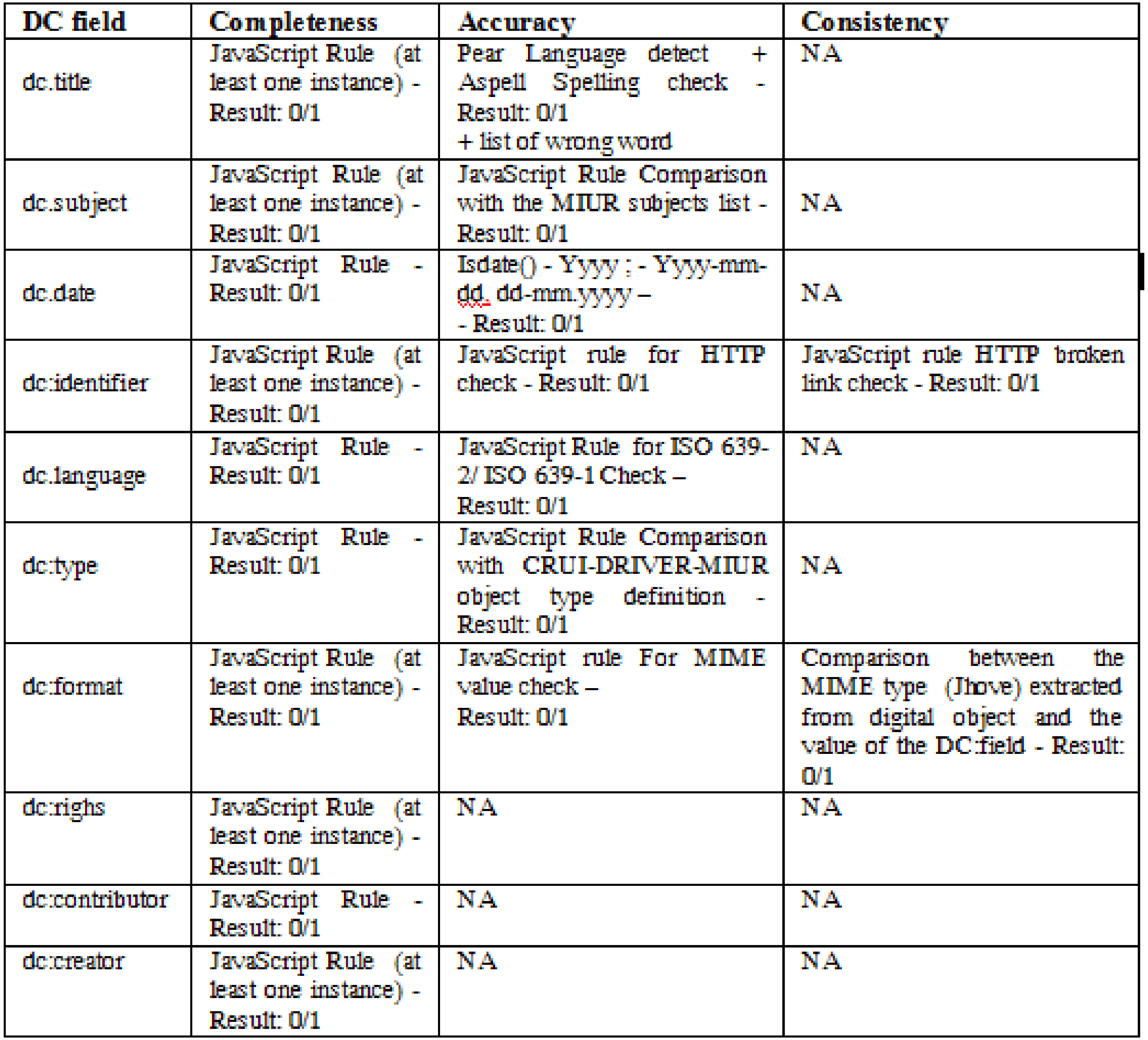
Completeness of a Field = $f(x) = \begin{cases} 0, \text{ if the field is empty} \\ 1, \text{ otherwise} \end{cases}$

Completeness of a Record *y* ComR(*y*)= $\dfrac{\sum_{i=1}^{nField(y)} f\,(x_i(y)) * w_i}{\sum_{j=1}^{nField_{Com}(y)} w_j}$ value ranged from 0 to 1.

Average Completeness of a Repository AvComR= $\dfrac{\sum_{i=1}^{n\mathrm{Re}cords(y)} ComR(y)}{n\mathrm{Re}cords}$ value ranged from 0 to 1.

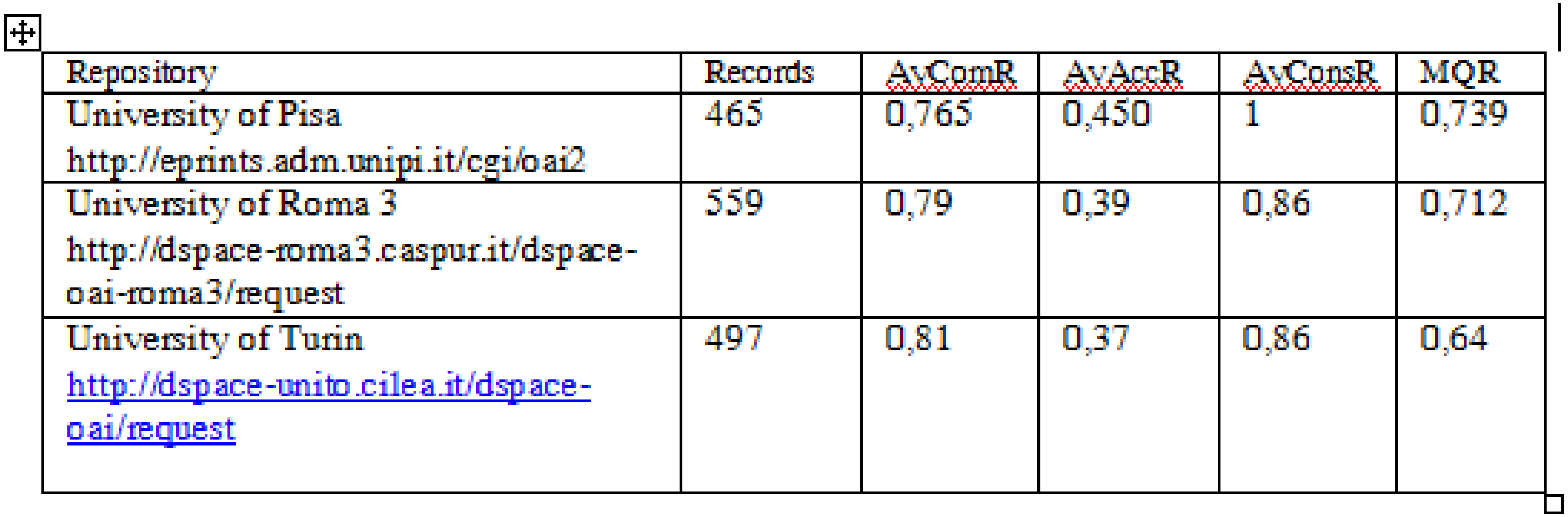
Quality of Repository *r* QR(***r***)= $\dfrac{AvComR(y)/\sigma^2_{Com} + AvAccR(y)/\sigma^2_{Acc} + AvConR(y)/\sigma^2_{Con}}{1/\sigma^2_{Com} + 1/\sigma^2_{Acc} + 1/\sigma^2_{Con}}$

Measurement methods



DC field	Completeness	Accuracy	Consistency
dc.title	JavaScript Rule (at least one instance) - Result: 0/1	Pear Language detect + Aspell Spelling check - Result: 0/1 + list of wrong word	NA
dc.subject	JavaScript Rule (at least one instance) - Result: 0/1	JavaScript Rule Comparison with the MIUR subjects list - Result: 0/1	NA
dc.date	JavaScript Rule - Result: 0/1	Isdate() - Yyyy ; - Yyyy-mm-dd. dd-mm.yyyy – - Result: 0/1	NA
dc:identifier	JavaScript Rule (at least one instance) - Result: 0/1	JavaScript rule for HTTP check - Result: 0/1	JavaScript rule HTTP broken link check - Result: 0/1
dc.language	JavaScript Rule - Result: 0/1	JavaScript Rule for ISO 639-2/ ISO 639-1 Check – Result: 0/1	NA
dc:type	JavaScript Rule - Result: 0/1	JavaScript Rule Comparison with CRUI-DRIVER-MIUR object type definition - Result: 0/1	NA
dc:format	JavaScript Rule (at least one instance) - Result: 0/1	JavaScript rule For MIME value check – Result: 0/1	Comparison between the MIME type (Jhove) extracted from digital object and the value of the DC:field - Result: 0/1
dc:righs	JavaScript Rule (at least one instance) - Result: 0/1	NA	NA
dc:contributor	JavaScript Rule - Result: 0/1	NA	NA
dc:creator	JavaScript Rule (at least one instance) - Result: 0/1	NA	NA

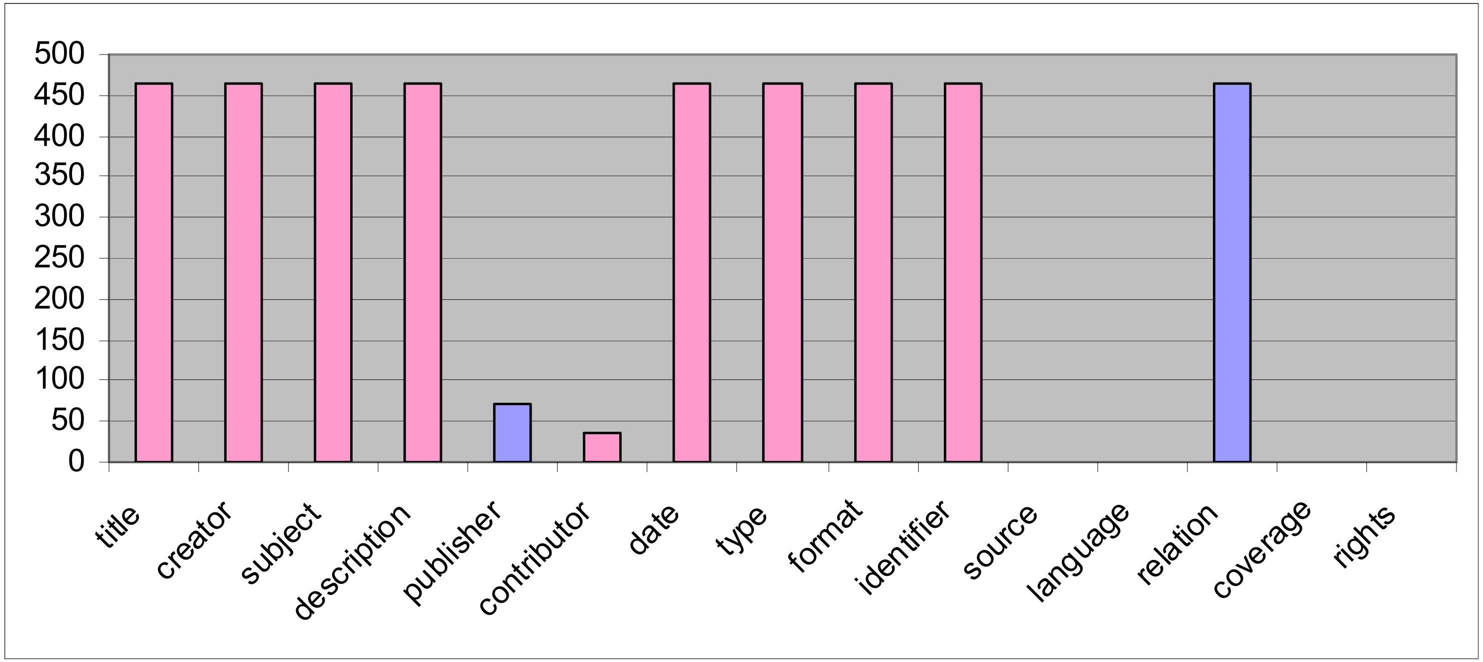
Assessment results examples



Repository	Records	AvComR	AvAccR	AvConsR	MQR
University of Pisa http://eprints.adm.unipi.it/cgi/oai2	465	0,765	0,450	1	0,739
University of Roma 3 http://dspace-roma3.caspur.it/dspace-oai-roma3/request	559	0,79	0,39	0,86	0,712
University of Turin http://dspace-unito.cilea.it/dspace-oai/request	497	0,81	0,37	0,86	0,64

Case study – University of Pisa in details

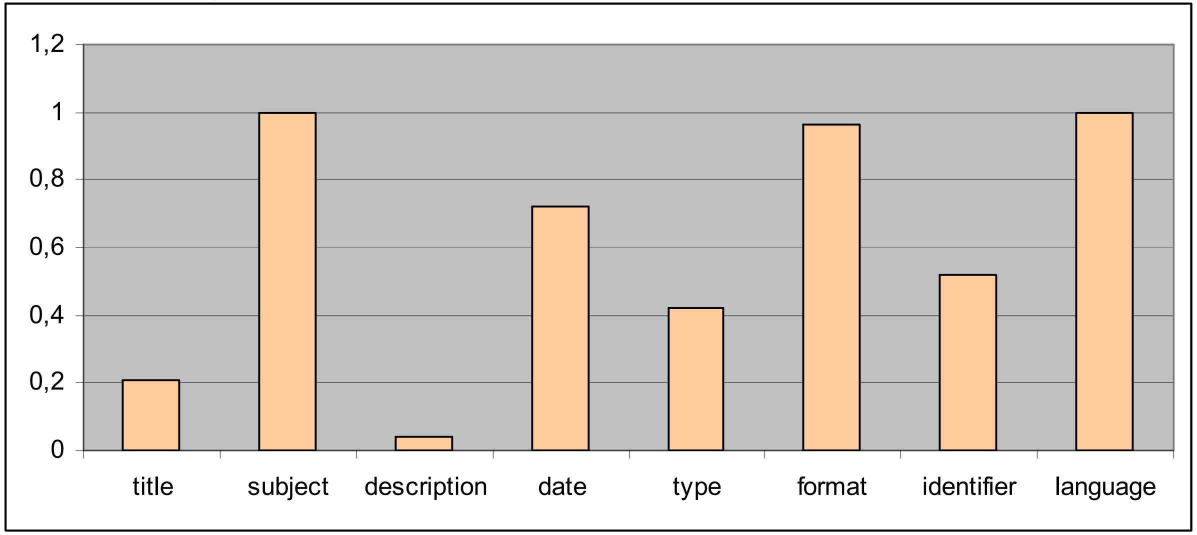
Completeness



Analysis: Only few records have the field Contributor with a value and no records have the field Language. This might means that a priori the repository system does not manage/ require those fields while for the others, their workflow seems reliable.

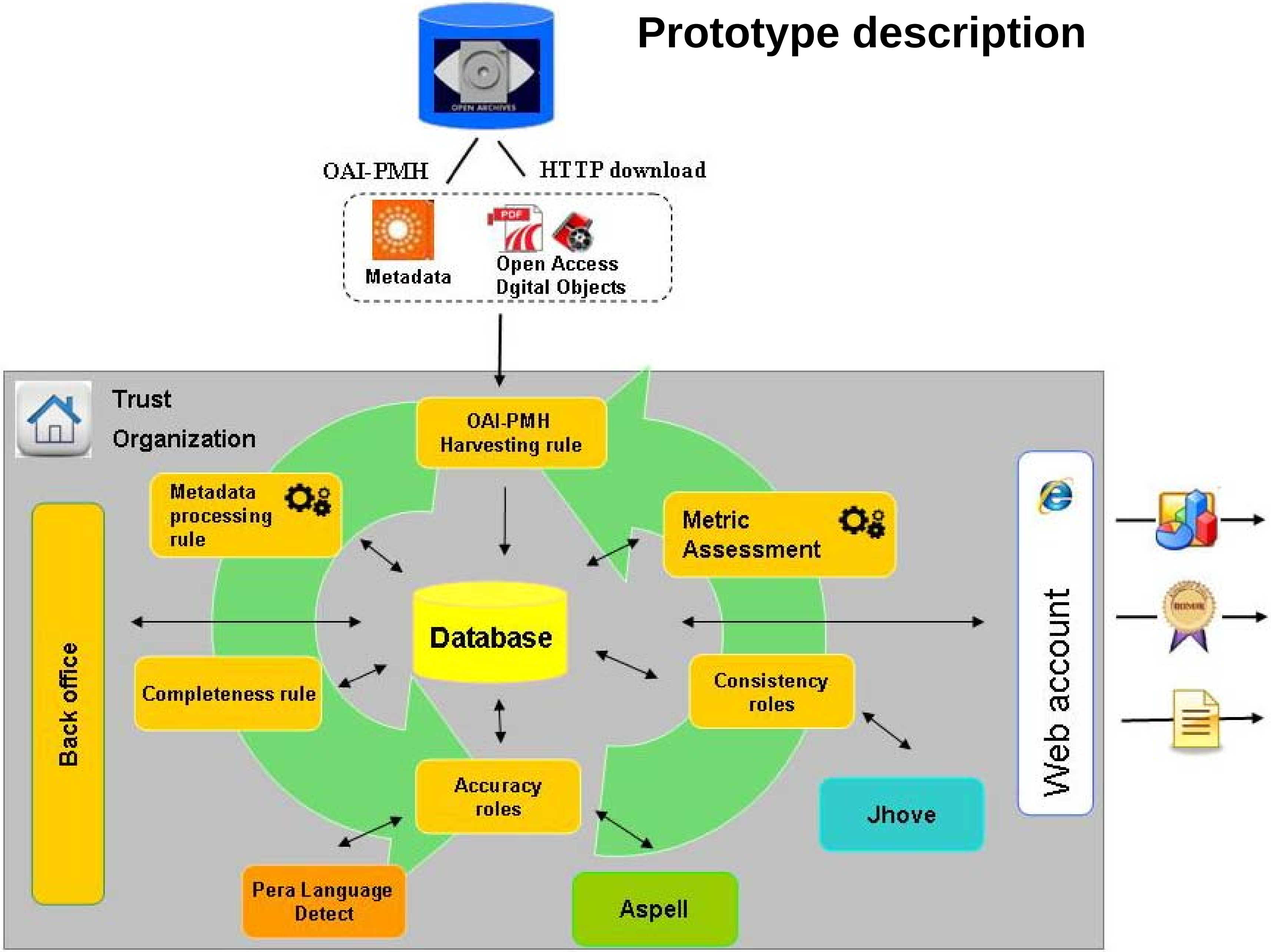
Case study – University of Pisa

Accuracy



Analysis: The fields Description and Title are those less accurate. This might be due to the type of the field (free text). Since the measurement criteria defined for those field are language detection and spelling check, this chart shows an high number of failures that might be due to typos for instance

Prototype description



Prototype description

Step 1: The process starts form the OAI-PMH harvesting form the Open Access repository.

The OAI-PMH harvester is implemented through an AXCP GRID rule. This process collects the metadata records and stores them in the database.

Step 2: The second step is performed by the metadata processing rule. This rule extracts each single field form the metadata table and populate a table with rdf-like tripe and each row represents a field.

Step 3: Then the rules for completeness assessment can be lunched.

Step 4: The accuracy can be assessed for each field through a proper evaluation rules, exploiting open source 3rd part applications like JHOVE or file format evaluation and ASPELL for spelling check.

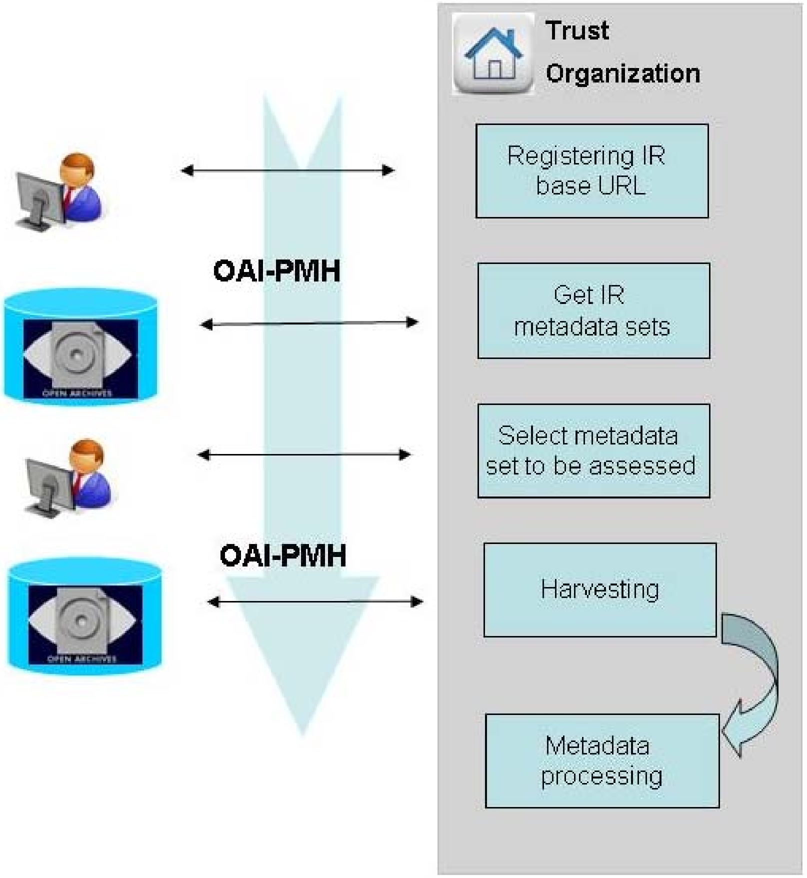
Step 5: This step addresses the consistency estimation. It can be lunched only on the field that have passed positively the completeness and the accuracy evaluation.

Step 6: The metric assessment, calculates MQ for the repository.

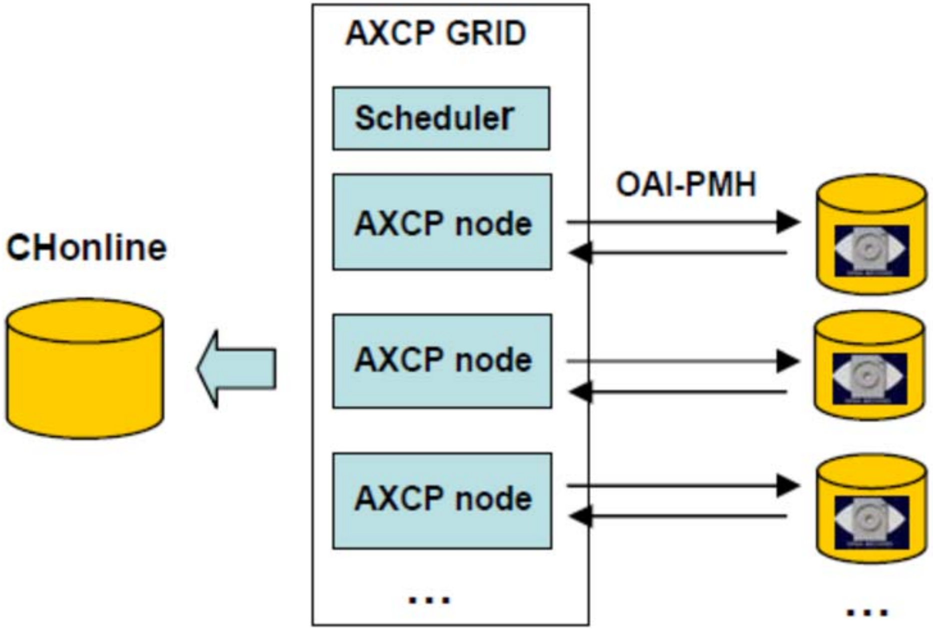
The prototype is based on AXMEDIS AXCP tool framework, an open source infrastructure that allows through parallel executions of processes (called rules) allocated on one or more computers/nodes, massive harvesting, metadata processing and evaluation, automatic periodic quality monitoring, and so forth.

Axmedis GRID infrastructure

Data collection workflow



Data collection through Axmedis GRID



3rd Party software

JHOVE - JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/

JHOVE provides functions to perform format-specific identification, validation, and characterization of digital objects. Format *identification* is the process of determining the format to which a digital object conforms. Format *validation* is the process of determining the level of compliance of a digital object to the specification for its purported format.

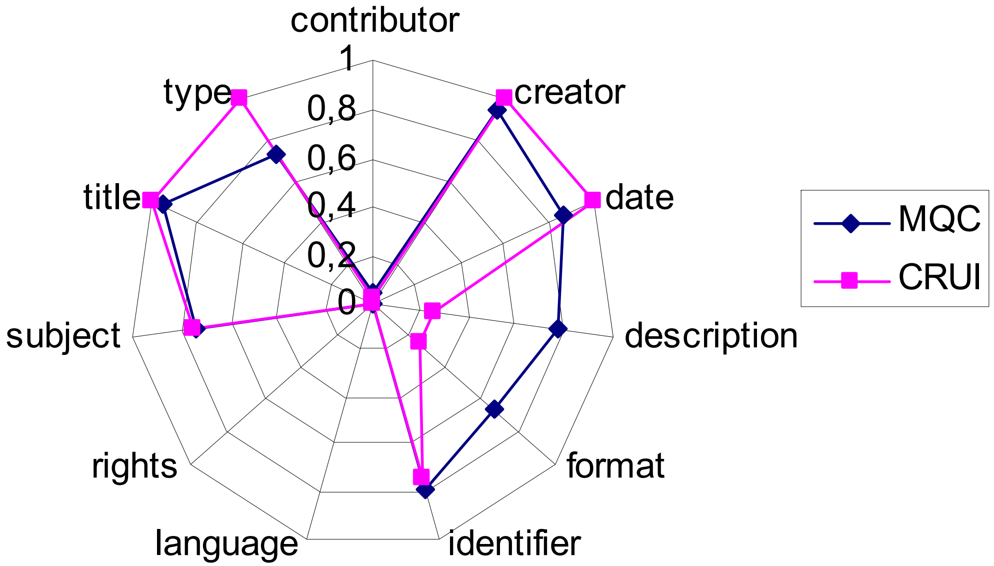
GNU ASPELL - http://aspell.net/

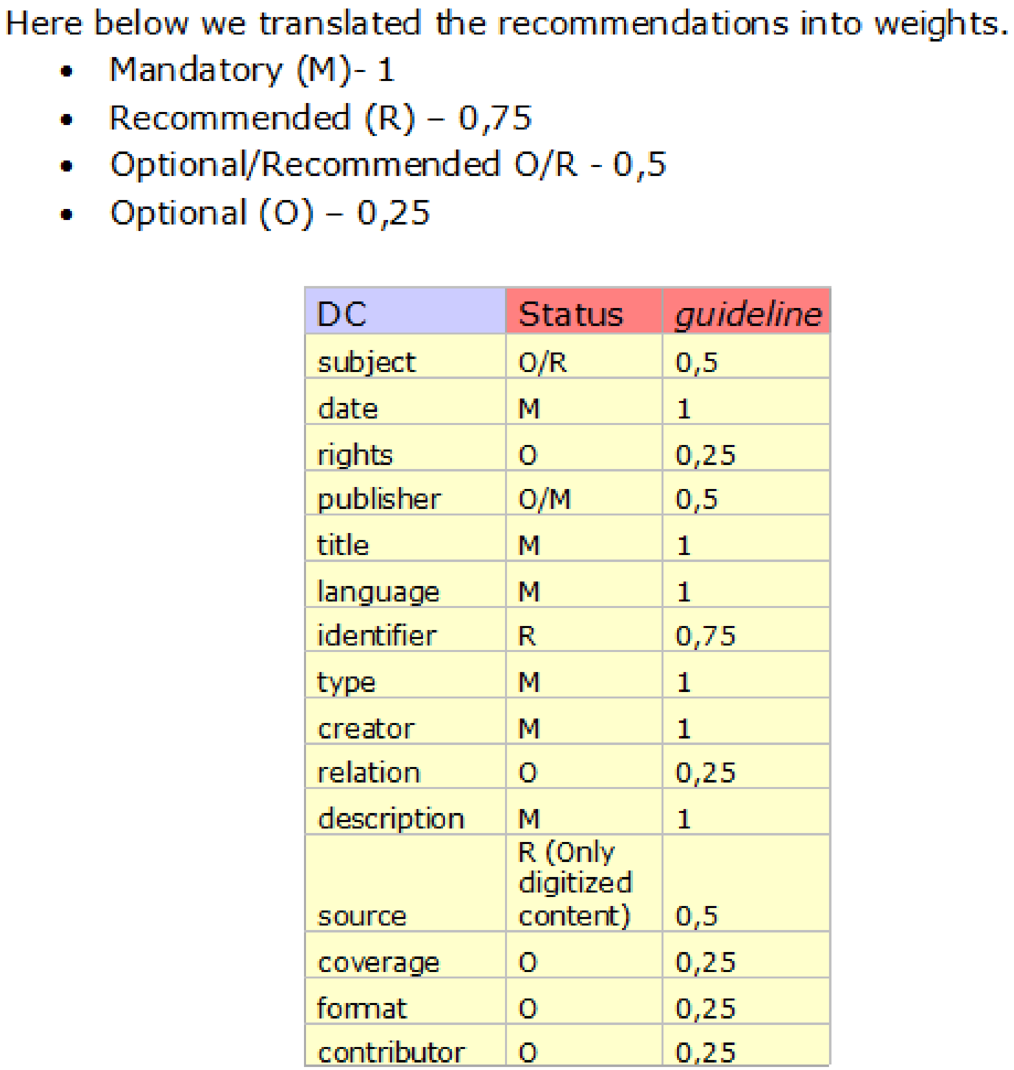
GNU Aspell is a Free and Open Source spell checker designed to eventually replace Ispell. It can either be used as a library or as an independent spell checker.

PEAR Language Detect http://pear.php.net/package/Text_LanguageDetect

The Per Language Detect is a Free PHP application able to recognize the language in input. The precision of the results depends from the length of the tens in input.

Comparison with other derived profiles





Here below we translated the recommendations into weights.

- Mandatory (M)- 1
- Recommended (R) – 0,75
- Optional/Recommended O/R - 0,5
- Optional (O) – 0,25

DC	Status	*guideline*
subject	O/R	0,5
date	M	1
rights	O	0,25
publisher	O/M	0,5
title	M	1
language	M	1
identifier	R	0,75
type	M	1
creator	M	1
relation	O	0,25
description	M	1
source	R (Only digitized content)	0,5
coverage	O	0,25
format	O	0,25
contributor	O	0,25

The Kiviatt chart shows the differences between our Quality Profile respect to Quality Profile derived from CRUI guidelines

Conclusions

a) The Completeness seems to be well addressed by all IR analyzed

b) There are some issues in the Accuracy dimension. The major problems were detected on the free – text fields such Title and Description

c) The DC is not expressive enough to support the complexity of the resources and their descriptive needs

d) We showed the validity of QP model respect to those derived from the other guidelines (e.g CRUI)

e) There are some cases in which the values could be considered accurate but their encoding format was not included in the our measurement model. Shared measuring modalities should be defined

Thank you very much!