Metadata Quality assessment tool for Open Access Cultural Heritage institutional repositories

Emanuele Bellini[1] , Paolo Nesi[1]

[1] Dept. di Ingegneria dell'Informazione
University of Florence – via Santa Marta ,
Florence, Italy
bellini@dsi-unifi.it, Paolo.Nesi@unifi.it

Abstract. Currently, the Metadata Quality in Cultural Heritage Institutional Repositories (IR) is an open issue. In fact, sometimes the value of the metadata fields contains typos, are out of standards, or are totally missing affecting the possibility of searching, discovering and obtaining the digital resource described. Scope of this work is to support institutions to assess the quality of their repository defining a Quality Profile for their metadata schema (e.g. Dublin core) and identifying the Completeness, Accuracy and Consistency as High level metrics. These metrics are translated in a number of computable Low level metrics (formulas) and measurement criteria. The quality measurement process has been implemented exploiting the Grid based AXMEDIS infrastructure to rise up the OAI-PMH harvesting and metadata processing performance. The quality profile metrics and the prototype have been tested on three Open Access Institution Repositories of Italian universities and the evaluation results are presented.

1 Introduction

The Metadata Quality (MQ) issue is still relatively unexplored, while there is a growing awareness of the essential role of MQ to exploit contents in the Cultural Heritage (CH) repositories. In fact, 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. While there is a wide consensus on the need to have high MQ, there are fewer consensuses on what high MQ means and much less in how it should be measured. Following the *Fitness for purpose* point of view, the [1] work considers high quality metadata if support the functional requirements of the system it is designed to support. In [1], internal and external functional requirements of metadata are defined in relation to the archive's web user interface such as search, browse, filter by, etc. These functional requirements are used to decide whose metadata are needed according to the aims of the archive, the designed community, the type of objects you are going to manage. In [2], the quality definition is related to the meeting or exceeding customer expectations or satisfying the users' needs and

preferences. Moreover, as stated in [3], the metadata relevance of a resource, and consequently their quality, has to be determined taking into account the *context of use.* For instance, a metadata record of absolute correctness and full completeness may not be of low quality because of the values of metadata fields do not comply with the context of use (domain standards and guidelines, e.g., wrong coding of language). Enforcing quality assurance during metadata creation [4] is one of the main concepts of the MQ. Thus, the semantic and descriptive elements associated with each resource in an institutional repository (IR), affect the quality of the service provided to the IR users. Similar to these approaches, that identifies the metadata requirements in relation to the final user expectations, are those presented in [5] and [6]. In [6], how the MQ affects the bibliographic function of research, use, dissemination, authenticity and management is described. The article defines that the main scopes of the metadata are related to retrieve, identify, select and deliver resources that are the main functions of online catalogues and digital libraries. In the Open Archive Information System (OAIS) standard [5], the Generate Descriptive Information (G-DI) function extracts Descriptive Information (DI) from the Archive Information Packages (AIPs) and collects DI from other sources to coordinate updates, and ultimately Data Management (DM). This approach includes metadata to support search and retrieval of Archive Information Packages (AIPs) (e.g., who, what, when, where, why). From the Library point of view, the QM reflects the degree to which the metadata performs the core FRBR functions of *find*, *identify*, *select* and *obtain* a digital resource [8].

In literature, the above mentioned functionalities, quality dimensions and metrics definitions are in general presented in a comprehensive Quality Features (QF). The QFs define several dimensions that the assessed information should comply in order to be considered of high quality. In [13], these QFs vary widely in their scope and goals. Some have been inspired by the Total Quality Management (TQM) paradigm, such as [14]; others are from the field of text document evaluation, especially of Web documents such as [15], others are linked to degree of usefulness or "fitness for use" [16] in a particular typified task/context. The NISO Framework of Guidance for Building Good Digital Collections presents six principles of what is termed "good" metadata [17]. These criteria and principles are defined by NISO to provide a framework of guidance for building robust digital collections, while they do not provide a clear number of well defined quality dimensions leaving the implementers free to address the issues in different ways. There are other metadata QFs that are formally defined and can be computed. They differ in granularity/detail, name of dimension, complexity and operational and there are many overlaps among them. In [18], three types of approaches to study information quality: 1) intuitive, 2) theoretical, and 3) an empirical approaches have been identified. The intuitive approach is identified when the researcher selects information quality attributes and dimensions using intuition and experience. In theoretical approach, quality features are a part of a larger theory of information/data relationship and dynamics, and, finally the empirical approach uses the information user data to determinate which dimension/feature the user applies for assessing information quality. In [19], 23 quality parameters are identified and some of them (e.g., ease of use, ease of creation, protocols, etc.) are more focused on the metadata schema standard or metadata generation tools. Stvilia in [20] uses most of them (excluding those not related with metadata quality), adds several more, and groups them in three dimensions of

Information Quality (IQ): Intrinsic IQ, Relational/Contextual IQ and Reputational IQ. The Stvilia's framework parameters includes accuracy, naturalness, precision, etc. Some of these parameters are grouped and included in comprehensive dimensions (completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility) by of Bruce & Hillman framework [21].

1.1 OA fragmented landscape

Several studies over the use of the metadata schema in Open Access repositories, such as those reported in [9], [10], [11], and [18], confirm the fragmentary landscape in terms of the interpretation of these schema, the policies adopted, the frequency of use of a certain field, and so forth. In [9], some criteria such as "Use of Metadata set" shown that the distribution of metadata set is quite spread and 153 different metadata schemas have been identified, over only 853 repositories; thus a high percentage. In general, there are several metadata standards promoted by different communities or even promoted and adopted by a single one. The most commonly spread are: Dublin Core[1](generally supported by default), METS[2], MPEG21 DIDL[3] (as a wrapper of other metadata models), MARCXML[4], etc. Unfortunately, there is a number of other sets such as Context_ob, Xepicur, junii, Uketd_dc that have been adopted by less than the 8% of the assessed archives. There exist a 15% of institutions using metadata sets which have been based on ad-hoc model (single instances in the distribution) or which do not have a significant number of institutions adopting them. The adoption of a non-standard metadata set and schema affects the effectiveness of archive visibility and distribution.

1.2 Issues in the schema implementations

When searching and browsing across archives, users expect to have typical search capacities also provided by single archive environment. The user will want to look for metadata records on documents that meet certain criteria, e.g., that belong to a certain author, or that date from a certain period of time. The language of the document might be relevant, or the user might be interested in documents that contain certain keywords in the title or abstract. In order to look for documents whose publication date might fall within a certain time period, the user should be able to formulate queries containing a comparison ("date before 2001-01-01 and date after 1999-12-31"). That implies that the dates contained in the metadata must be well formed, computable and comparable, there must be a uniform date format and an ordering on that format. In [9] some problems regarding the interpretation and use of the single metadata fields have been detected. Moreover, it is well know that the use of simple

[1] http://www.openarchives.org/OAI/2.0/oai_dc.xsd
[2] http://www.loc.gov/standards/mets/mets.xsd
[3] http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd
[4] http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd

Dublin Core foresees a high level of flexibility for filling in the metadata field. In fact analysis as [9], has shown that a very few number of institutions adopted a qualified DC model, as defined by standard recommended best practice with a controlled vocabulary such as RFC 4646 or ISO639-1. For example, when the user is looking for an author, he is not interested in other information, thus, if the author field contains address and affiliation the system should distinguish between the author name and the rest of the information [12].
Moreover, the metadata multi-language system is managed in two modalities: using different instances of DC fields for each language or expressing different languages in the same field with a separator. The analysis has outlined that this separator can be arbitrary as: ‘ ,’ ‘ ;’, ‘ – ‘, ‘/’. The instance had value like en, eng, English, en_GB, en-GB, for English or es, spa, Espanol, Spanish; , spa; sp for Spanish and so forth.
Regarding the DC:format field, in [9] different filling modalities have been found with the presence of the file format definitions, physical medium descriptions, the dimensions of the resource and as described by standard definition while the recommended best practice refers to use a controlled vocabulary such as the list of Internet Media Types.
The scope of this work is to support institutions to obtain higher level of MQ for their repository trough a continue or sporadic quality assessment. Thanks to the low effort required for the assessment (automatic) and the scalability of the technology infrastructure adopted, the proposed solution is particularly suited for the institutions with low resources to manage and review the related metadata. Therefore, the main goal of this work is to assess the MQ to support cultural heritage institutions in obtaining and maintaining an appropriate quality level of their IR in a very simples and economical way, defining: (a) a MQ Profile and related dimension able to be assessed through automatic processes, (b) s set of metrics to be used for assessing and monitoring MQ, (c) a technological tool to asses the metrics defined based on a scalable infrastructure, thus estimating reference values from the global state of the quality.
The article is divided in the following sections: in section 2, the MQ Framework with the definition of quality profile, and metric dimensions, formulas and the assessment results on three IR of Italian universities are provided; in section 3, the prototype and its features are described; conclusions are reported in section 4.

2 Metadata Quality Framework

In order to address that transparency and objectivity required for a quality assessment, the adoption of a standard methodology for design metrics and manage the entire workflow is crucial. Although Goal Question Metric originated as a measurement methodology for software development, the basic concepts of GQM can be used anywhere than effective metrics are needed to assess satisfaction of goals [22]. The literature typically describes GQM in terms of a six-step process while in [22] these 6 steps are compressed in the following four phases that this work has adopted as a basis of the entire research workflow:

a) Planning Phase: This phase is represented by the Metadata Quality Profile definitions. (MQP) As stated, the MQP is based on the goal or purpose of metadata records into the OA domain and drives the metrics definition.
b) Definition phase: The definition phase consists in defining the High Level Metrics (HLM) according to the MQP and through the GQM top-down approach, the Low Level Metrics (LLM).
c) Data collection phase: Once metrics are identified, one can determine what data items are needed to support those metrics, and how those items will be collected. A Measurement Plan is defined according to [27] and includes: the definitions of direct measurements with all possible outcomes (values), the medium (tools) that should be used for collecting the measurement, and the definition of derived measurement.
d) Interpretation phase: The last step of GQM process is about looking 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.

2.1 Metadata Quality profile

As we stated before, every quality assessment requires a definition of a clear and stable baseline quality of reference in a given context, called **Quality Profile (QP)**. A QP allows of taking into account the user perspective in the definition of the baseline quality of reference. The QP has to reflect also the notion of the quality of the OA user community and it is worth to notice that a QP must be agreed among all stakeholders involved in the quality assessment. Thus, in order to address this requirement, we submitted a specific questionnaire to the Open Access community with the aim of gathering their points of view about relevance of each DC field in a DC records quality assessment.

Data Filtering
In order to get more confident results from the analysis, we filtered out the answers with the following criteria: Critical target (we have focused Researchers 20,6%, Professors 12,7%, ICT experts15,9%, Archivists15,9%, Librarians 25,4% avoiding not relevant data provided by user with not suitable profiles that may introduce "noise" in the statistics analysis), level of knowledge (the 17% of the responders stated their knowledge of the DC schema is less then 5 in a range from 1 to 10), never worked with metadata (the work 6,3% of the responders does not include the definition and use of metadata), and never dealt with metadata quality (the 11,1% of the responders has never dealt with the quality of metadata). Then we calculated the Average, Variance and the level of confidence from the answers for each DC field before and after the data filtering. The results has shown a reduction of the Variance for each field after the data filtering as a confirmation of the correctness of our assumption.

Field selection
In order to define the quality profile, weaimed to determinate which are the fields to be taken in more consideration. In fact, each field has a different level of relevance in

a record. The relevance has been estimated asking to the Open Access community experts to assign a relevance to each DC field 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). Thus we defined the following criteria to exclude those fields that are not considered relevant by the OA community, from the quality assessment:
- The quality assessment on the field *f* can be avoided if the Average weight is 5,5 or less;
- The quality assessment on the field *f* can be avoided if the difference between the Average weights and ½ of the level of confidence is 5,5 or less.

According to the field selection criteria defined, the results show that Coverage, Publisher, Relation and Source have not passed the threshold of 5.5. In fact, the Average of the Source field score is under the threshold (5.119) yet, while for the other fields the differences between the Average and the relative level of confidence are Coverage: 5.334, Publisher: 5.325, Source:4.923 respectively. This assessment allowed us to identify relevant fields to be taken into account in the evaluation of the quality assessment. The relevance weights assigned to each field are the normalized Averages of the weights assigned by the AO experts (see Table 1).

Table 1 MQ Profile, relevance weights

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

2.2 High Level metrics definition

The MQ dimensions provided can be assessed at three levels: metadata field, metadata record, and repository level. In particular, the metadata field level foresees metrics that are able to evaluate the Completeness, Accuracy and Completeness for each metadata field defined by the schema. The derived measures give quality indexes on the fields' implementation into the repository. The metadata record level foresees metrics that, compounding the field metrics properly, are able to evaluate the quality dimensions at record level. The derived measures give quality indexes for the total amount of the Metadata records managed by a repository. The third level foresees a clustering of the quality results obtained from the first and/or the second level to provide an overview of the repository metadata quality. To this end, Consistency

evaluation can be performed only if the Accuracy evaluation is passed. The Accuracy can be assessed in the Completeness evaluation is successfully passed.

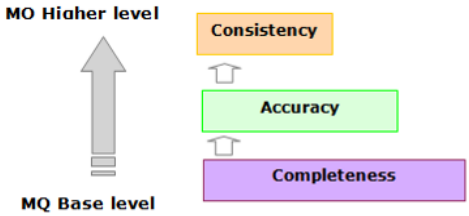

Figure 2 – Multi level MQ assessment

Hence, the Base level of Metadata Quality is assured by the full completeness of the metadata fields in the IR. Built upon this result, the Accuracy assessment can be performed. The Accuracy box is smaller the Completeness one because the number of field analyzed in this process is less than the number of fields assessed during the Completeness evaluation, where all metadata fields are taken into account. The same consideration is for Consistency box respect to the Accuracy one. This is due to the fact that for some fields is really difficult to evaluate accuracy or consistency dimension with an automatic process.

2.2.1 Completeness

Commonly the concept of Completeness is related to the presence of uncompleted fields in a record, and can be generically defined as the degree to which values are available with respect to the required [24]. In [21] instead, the Completeness does not mean that all the metadata elements are used in a given metadata schema because of two main reasons: "*First, the element set used should describe the target objects as completely as economically feasible.[…] Second, the element set should be applied to the target object population as completely as possible.*" It is clear that, there are different ways of considering complete a metadata record by a user or by a community.

Unfortunately, this approach does not seem to be feasible for a certification purpose because of its variability and uncertainty along the time. In fact, if some fields are usually not filled, it does not mean that they are not required or needed. There are several reasons that can determinate an empty value in a field. In [1], analyzing the quality of metadata in an eprint archive, the authors have identified in the publication workflow and eprint software customization the main issues. In summary, the Completeness dimension is function of the relevance weights assigned to the field by the Designed Community[5] according to recognized standards and guidelines.

[5] Designated Community: An identified group of potential Consumers who should be able to understand a particular set of information. The Designated Community may be composed of multiple user communities – IS0:14721:2003 OAIS Reference Model

2.2.2 Accuracy

In the Bruce and Hillman framework [21], the metadata should be accurate in the sense of high quality editing, thus we consider accurate a record when:

- there are not typographical errors in the free text fields,
- the values in the fields are in the format expected.

The same point of view is adopted by Stvilia [18], when defines Accuracy/Validity dimension of the Intrinsic IQ as: "*the extent to which the content information is legitimated or valid according to some stable reference source such as a dictionary, standard schema and/or set of domain constraints and norms*". As an example, the Accuracy evaluation can be performed taking into account recommendations such as the use of ISO639-1 standard for the DC:language. Again, in the CRUI Metadata Working Group report, it is specified that the DC:subject has to assume the MIUR disciplinary sector values, while the DC:type field value has to be compliant with the MIME[25] definition, where an URI 6 is expected (DC:identifier), thus, a syntax correctness check is required. In summary, there is an Accuracy issue when a metadata record includes values not defined in the standards. Indeed, the Accuracy (correctness) could be a binary value, either "right" or "wrong", for objective information like file type, language, typos, and so with respect to the values expected by the standard.

2.2.3 Consistency

Some synonyms of Consistency referred to the metadata can be: compliance, non-contradictory, and data reliability. From our research perspective, the Consistency dimension has to address the logical error. In a metadata record, the results of a missed consistency control can affect several fields. Examples are:

- a resource results "published" before to be "created" (data fields), the MIME type declared is different respect to the real bitstream associated,
- the language of the Title is different respect to the object description, and
- the link to the digital objects is broken.

Some of the Consistency cases are difficult to be detected automatically or required notable computing efforts. For instance, the assessment of the MIME type can be performed only if the resource is downloaded and processed and a strong scalable infrastructure is required. The consistency issue affects another crucial field in a metadata schema like the fields used to obtain the resource, for example via URL. In this case, the consistency issue is related to the actual access to the resource. In general, this issue occurs when the URL to the resource is for instance, a broken link. This can happen for different reasons such as the digital object is moved to another server and the link has not been updated or the URL is written in a wrong way, and so forth. In this sense, the consistency assessment on those fields is based on the check of the effective access to the content file. In summary, the consistency issues emerge when the value in the field is formally compliance to the standard but is logically wrong.

---

[6] Uniform Resource Identifiers IEFT RFC 3986

2.3 Metric implementation

The overall approach and aim of this work reflect the measurement objectives proposed in [26]. In order to avoid the risk of getting overwhelmed with data, as outlined in [27] and [28] one factor of defining successful measurement frameworks is to start with the most important measurements and grow slowly as the organization matures, especially if measurements are being tried for the first time. Thus, the basic measures of the three dimensions are applied on each single field and are represented by the following functions*:*

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

Accuracy of the field is defined by $g(x)=\begin{cases}0, \text{if an accuracy issue was detected}\\ 1, \text{no problem founded}\end{cases}$

Consistency of the field is defined by $h(x)=\begin{cases}0, \text{if a consistency issue was detected}\\ 1, \text{no problem founded}\end{cases}$

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

Accuracy of a Record *y* AccR(*y*)= $\frac{\sum_{i=1}^{nField_{Acc}(y)} g(x_i(y))*w_i}{\sum_{j=1}^{nField_{Acc}(y)} w_j}$ value ranged from 0 to 1.

Consistency of a Record *y* ConR(*y*)= $\frac{\sum_{i=1}^{nField_{Con}(y)} h(x_i(y))*w_i}{\sum_{j=1}^{nField_{Con}(y)} w_j}$ value ranged from 0 to 1.

The derived measures for the quality are:
the average (Av)of the quality score for each dimension

AvComR= $\frac{\sum_{i=1}^{n\text{Re}cords(y)} ComR(y)}{n\text{Re}cords}$; AvAccR= $\frac{\sum_{i=1}^{n\text{Re}cords(y)} AccR(y)}{n\text{Re}cords}$; AvConR= $\frac{\sum_{i=1}^{n\text{Re}cords(y)} ConR(y)}{n\text{Re}cords}$;

Mean Quality of Repository *r* QR(***r***)= $\frac{MComR(y)/\sigma_{Com}^2 + MAccR(y)/\sigma_{Acc}^2 + MConR(y)/\sigma_{Con}^2}{1/\sigma_{Com}^2 + 1/\sigma_{Acc}^2 + 1/\sigma_{Con}^2}$
value ranged from 0 to 1.

where:
x is the i-th field in the schema; *y* is the *y*-th record; $nField_{Com}$: the total amount of fields in the metadata schema selected for the completeness evaluation, $nField_{Acc}$ the total amount of metadata fields selected for the accuracy evaluation, $nField_{Con}$ the total amount of metadata fields selected for the consistency inspection, and *nRecord*(r)is the number of records in the IR, r.

The table below reports the main measurement criteria to assess the quality dimensions for each DC field. In particular, for the accuracy and consistency dimensions 3rd party tools are used for language recognition, spelling check and MIME type extraction.

Table 2 Measurement criteria

DC field	Completeness	Accuracy	Consistency
dc.title	Javscript Rule (at least one instance) - Result: 0/1	Pear Language detect + Aspell Spelling check - Result: 0/1 + list of wrong word	NA
dc.subject	Javscript Rule (at least one instance) - Result: 0/1	Javascript Rule Comparison with the MIUR subjects list - Result: 0/1	NA
dc.date	Javacript Rule - Result: 0/1	Isdate() - Yyyy ; - Yyyy-mm-dd - dd-mm.yyyy - Result: 0/1	NA
dc:identifier	Javscript Rule (at least one instance) - Result: 0/1	Javascript rule for HTTP validator - Result: 0/1	Javascript rule HTTP broken link check - Result: 0/1
dc.language	Javscript Rule - Result: 0/1	Javascript Rule for ISO 639-2/ ISO 639-1 Check - Result: 0/1	NA
dc:type	Javscript Rule - Result: 0/1	Javascript Rule Comparison with CRUI-DRIVER-MIUR object type definition - Result: 0/1	NA
dc:format	Javscript 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	Javscript Rule (at least one instance) - Result: 0/1	NA	NA
Cd:contributor	Javscript Rule - Result: 0/1	NA	NA
Cd:creator	Javscript Rule (at least one instance) - Result: 0/1	NA	NA

2.4 Assessment Results

The table 3 shows the results of a quality assessment conducted on 3 Open Access Institutional Repositories of Italian universities. It is worth to notice that the completeness and consistency indicators thand to obtain an high value whiale Accuracy tends to be less of 0.5, thus a low quality. The free text fields like description or language tends to be critical because of the presence of typos or not standard compliant values

Table 3 Assessment results

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

3 System Architecture

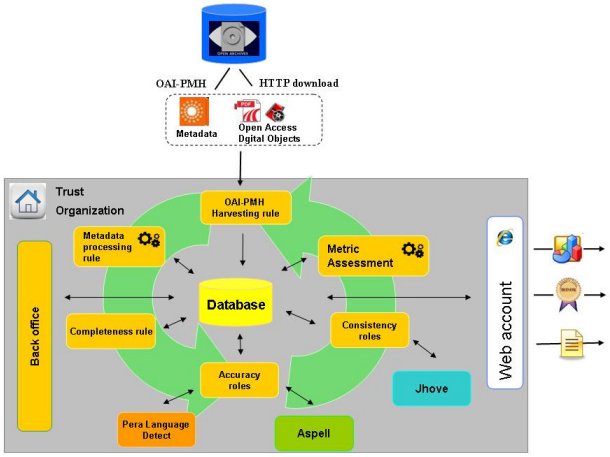

Figure 1 System Architecture

The MQ assessment tool implements a number of GRID rules that identify the steps of the quality assessment:
Step 1: The process starts form the OAI-PMH [29] harvesting form the Open Access repository. The OAI-PMH [29] 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 rule.
Step 5: This step addressesthe 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.
This MQ service is based on AXMEDIS AXCP tool framework, an open source infrastructure that allows trhough 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 [7].

The rules are managed by a central scheduler and are formalized in extended JavaScript [30]. The AXCP Scheduler performs the rule firing, node discovering, error report and management, fail over, etc. The scheduler may puts rules in execution

(with parameters) periodically or when some other application request. It provides reporting information (e.g., notifications, exceptions, logs, etc…) to external workflow and tools by means of WEB services. The control and activation of rules can be performed via a Web Service through the Rule Scheduler, by any program and web applications, for example workflow tools (systems such as Open Flow and BizTalk), PHP, CGI, JSP, etc. The single node could invoke the execution of other rules by sending a request to the scheduler, so as to divide a complex rule into sub-rules running in parallel and use the computational resources accessible on the grid. An AXCP rule may perform activities of content and metadata ingestion, query and retrieval, storage, semantic computing, content formatting and adaptation, extraction of descriptors, transcoding, synchronisation, estimation of fingerprint, watermarking, indexing, summarization, metadata manipulation and mapping, packaging, protection and licensing, publication and distribution. AXCP nodes have plug-ins or may invoke external tools to expand capability with customized/external algorithms and tools.

3.1 GRID based metadata harvesting and processing

The solution approach is based on OAI-PMH protocol, a REST-based full Web Service that exploits the HTTP protocol to communicate among computers, using either the GET or the POST methods for sending requests. According to OAI-PMH protocol, Guidelines for Harvesting Implements [29] and OA implementation tutorial, a client may put a request to OAI server to ask for the stored content descriptors. Answers are related to the accessible records, and adopted formats. The OAI-PMH protocol provides a list of discrete entities (metadata records) by XML stream. As it occurs with a web crawler, the harvester contacts and inspects the OA data providers automatically and it extracts metadata sets associated with digital objects via OAI-PMH protocol. Because of the computational weight of these processes, the harvester has been implemented by using the grid based parallel processing on DISIT cloud computing infrastructure. The grid solution has been realized by using AXMEDIS Content Processing (AXCP GRID). The computational solution has been implemented by realizing a parallel processing algorithm written in AXCP Extended JavaScript [30]. The algorithm has been allocated as a set of periodic processes replicated on a number of grid nodes, typically from 1 to 15 max. The process is managed by the AXCP Scheduler. It is possible to put in execution a number of rules that are distributed to the available grid nodes. Each rule is a ‘harvester’ executor of an OAI-PMH request to obtain the metadata records, parsing the XML response and storing information in our local database. This solution reduces the computational time up to a factor equal to the number of nodes used for completing the harvesting of repositories. In effect, the parallel solution is not only an advantage for the speed up, but also for the reduction of the time needed to get a new global version of the metadata collected in the OI repositories.
The metadata harvesting is the first step to collect data and per se it is not sufficient to evaluate the quality of metadata implementation thus an additional grid rule got the XML of each non processed record stored in the database and it extracted the single fields. Therefore, each field of each specific record has been stored with its value, type, and additional information in the database. This poses the basis to perform a

deeper analysis, as described in the following. This process led to a sort of an extended RDF model and thus to a metadata normalization allowing queries on the single fields. This table turned out to be very huge (for each field of each metadata record a detailed field record is generated. For instance 15 new records are generated from a single DC based metadata record). The resulting table of single fields has been mainly used as a metadata assessment for the purpose of this work.

4 Conclusions

The MQ issue can be addressed though automatic tools if it is possible to identify a number of criteria that can be computable and comparable against a baseline of quality. Sometimes this MQ of reference defined by user community is not strictly aligned with the official standards and guidelines, thus a specific quality profile should be identified to be more effective in the MQ assessment. This work, following the GQM approach has defined a) a MQ profile for CH repositories, b) identified three High Level Metric and their related formulas taking in to account their computability, c) implemented a measurement strategies exploiting 3rd parties and original software roles d) integrated all these components in an online cost-effective service to support Cultural heritage institution in maintaining high MQ in their repositories.

References

1. Guy M., Powell A. and M Day "Improving the Quality of Metadata in Eprint Archive" ARIADNE Issue 38 Jenuary 2004 http://www.ariadne.ac.uk/issue38/guy
2. Evans & Lindsay – The management of quality control (6 ed.) Mason, OH Thompson
3. Margaritopoulos T., Margaritopoulos M., Mavridis I., Manitsaris A., A conceptual framework for metadata quality assessment, Proceeding DCMI '08 Proceedings of the 2008 International Conference on Dublin Core and Metadata Application
4.Barton Jane, Currier Sarah, Hey Jessie M.N., Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice, in Proceedings 2003 Dublin Core conference: Seattle, Washington, USA, 28 September - 2 October 2003, (NY): Information Institute of Syracuse, 2003, p. 39-48,
5.ISO 14721:2003 Reference Model for an Open Archival Information System (OAIS)
6.Park Jung-Ran, Metadata quality in digital repositories: A survey of the current state of the art, "Cataloging & classification quarterly", vol. 47, nos. 3-4 (April 2009), p. 213-228
7. IEEE Multimedia: P. Bellini, I. Bruno, D. Cenni, P. Nesi, "Micro grids for scalable media computing and intelligence on distributed scenarious", IEEE Multimedia, Feb 2012, Vol.19, N.2, pp.69.79, IEEE Computer Soc. Press
8.IFLA - Functional Requirements for Bibliographic Records FRBR, Final Report http://archive.ifla.org/VII/s13/frbr/frbr1.htm. 1998
9.Bellini Emanuele, Deussom Marcel Aime, Nesi Paolo, Assessing Open Archive OAI-PMH implementations – DMS2010
10. Bui, Yen; Park, Jung-ran An assessment of metadata quality: A case study of the National Science Digital Library Metadata Repository , IST Research Day 2006 11. Efron Miles: Metadata Use in OAI-Compliant Institutional Repositories.J.Digit.Inf.8(2): (2007)

12. Fischer Gudrun, Fuhr Norbert - Heterogeneity in Open Archives Metadata-Cyclades project
13. Ochoa Xavier and Duval Erik: Towards Automatic Evaluation of Metadata Quality in Digital Repositories http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.89.8371
14.Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Communications of the ACM 40(5) (1997) 103–110
15.Zhu, X., Gauch, S.: Incorporating quality metrics in centralized/distributed information retrieval on the world wide web. In: Research and Development in Information Retrieval. (2000) 288–295
16.Jura, J - Juran on quality by design - New York, NY 1992 Free press
17.NISO A Framework of Guidance for Building Good Digital Collections (Bethesda, MD:NISO Press, 2007),61-2
18.Stvilia – Measuring Information Quality Dissertation – Urbana- Illinois, 2006
19. Moen, W.E., Stewart, E.L., McClure, C.R.: Assessing metadata quality: Findings and methodological considerations from an evaluation of the u.s. government information locator service (gils). In: T.R. Smith (ed.) ADL '98: Proceedings of the Advances in Digital Libraries Conference, pp. 246–255
20.Stvilia, B., Gasser, L., Twidale, M.: A framework for information quality assessment. Journal of the American Society for Information Science and Technology 58(12), 1720–1733 (2007)
21.Bruce, T.R., Hillmann, D.: Metadata in Practice, chap. The continuum of metadata quality: defining, expressing, exploiting, pp. 238–256. ALA Editions, Chicago, IL (2004)
22.Basili V.R, Caldiera G. and Rombach H. D., "The goal question metric approach", in Encynclopedia of Software Engineering. Wiley, 1994
23. Van Solingen Rini and Berghout Egon - The Goal/question/Metric Method: a pratical guie for quality improvment of software development – The Mcgraw-hill companies ISBN0077095537
24.Pipriani Baba and Ernst Denise– A Model for Data Quality Assessment
25.IETF RFC 2045 Multipurpose Internet Mail Extensions (MIME), 1996
26.Zubrow David - Can you trust your data? Measurement and Analysis Infrastructure Diagnosis 2007 SEI
27.ISO/IEC IS 15939, Software Engineering – Software Measurement Process, 2002
28.Berander Patrik, Jönsson Per - A Goal Question Metric Based Approach for Efficient Measurement Framework Definition International Symposium of Empirical Software Engineering (ISESE '06) Rio de janeiro Brazil
29.Lagoze Carl, Herbert, Michael Nelson, 2002. Implementing Guidelines for the Open Archives Initiative for Metadata Harvesting: Guidelines for Harvesting Implementes
30. Bellini P., Bruno I., Nesi P., "Visual Programming of Content Processing Grid", The 15th International Conference on Distributed Multimedia Systems, DMS2009.