Ph.D. program in

"Telematica e Società dell'Informazione"

Ciclo XXIII

Instituted by the Italian University Consortium among

University of Florence

University of Siena

METADATA QUALITY CERTIFICATION SERVICE FOR OPEN ACCESS INSTITUTIONAL REPOSITORIES

A thesis submitted for the degree of
Doctorate of Philosophy

Candidate: Emanuele Bellini

Coordinator: Prof. Dino Giuli

Supervisors:Prof. Paolo Nesi

CODE :

DATE

Content

Introduction

The Web has drastically changed the information environment where users of the humanities work and study and the information needs to be more accessible to become collective knowledge. These changes are affecting also the scientific domain where new technologies allows new dissemination and exploitation opportunities of research product.
The Declaration on Open Access, OA, to Knowledge in the Sciences and Humanities[1] aims to support these new opportunities asserting that " Our mission of disseminating knowledge is only half complete if the information is not made widely and readily available to society. New possibilities of knowledge dissemination not only through the classical form but also and increasingly through the open access paradigm via the Internet have to be supported."

So far, according to this declaration, a lot of cultural heritage and scientific institutions are implementing open access institutional repositories. The actual Open Access implementation landscape is really fragmented and some difficulties prevent its wide adoption and exploitation. For instance there are some disciplines such as medicine or engineering that are slow in adopting the Open access paradigm while in the physicians community is a common an well accepted practice.
In particular, the Open access contributions must satisfy two conditions [2]:

1. The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship (community standards, will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use.

[1] http://oa.mpg.de/berlin-prozess/berliner-erklarung/
[2] http://oa.mpg.de/berlin-prozess/berliner-erklarung/

2. A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well established organization that seeks to enable open access, unrestricted distribution, inter operability, and long-term archiving.

Even if the principles are corrects, the current world wide Open access repository implementations are very fragmented in term of contents managed, metadata, level of service provided, etc.

There are several studies such as [Bellini, Deussom, Nesi,2010] that shows these differences. Some initiatives are started as an experiment managed by excited volunteers affiliated to the University Library department but sometimes they are not evolved into a stable applications because of a lack of policies, workflow an responsibilities definitions.

For instance, more than 70 Italian universities have signed the Berlin Declaration, but there are difficulties in inserting officially the Open access declaration in the institution statutes. Moreover, these difficulties and delays prevent the use of the Open access resources in the national Research Evaluation performed in Italy by ANVUR.

There is not a general agreement about terminology as 'deposit', 'archive', 'repository', etc because each terms is related to a particular objective.

In fact, in the "Pathfinder Research on Web-based Repositories" article [Ware, 2004], Mark Ware outlines that the objectives of institutional repositories can be very diverse. In this work we adopt the term "Institutional Repository" (IR) for referring to open access institutional repositories implemented in the universities and research institutions with the objectives to collect, organize and disseminate open access scientific resources and their metadata, in order to contribute to improving the research results visibility [Foulonneau, André]. We cut the "preservation" objective from the list of the because, the prevalent orientation of the institution now, it to delegate the preservation strategies of thier resources to external service such as the national legal deposit.

These repositories can contain a wide rage of scholarly publications (reports, working papers, pre- and post-prints of articles and books

of research institutions, etc.) produced by research institutions. In any case, they contribute to the Open Access movement by providing platforms for researchers to make research results such as papers or technical reports freely available on the web.
In order to address this aim, a number of software tools is available for implementing an open institutional repository such as Dspace [3], Fedora [4], Eprints[5], Greenstone [6], etc. These software are, in general, OAIS Standard oriented [ISO 14721, 2003] open source and implement the OAI-PMH protocol.
The Open Archive Initiative has developed the OAI-PMH protocol for publishing and thus making possible the metadata harvesting among repositories. The OAI architecture identifies two logical roles: "Data Providers" and "Service Providers". Data Providers deal with both the deposit and the publication of resources in a repository they "expose" to provide the metadata about resources in the repository. They are the creators and keepers of the metadata and repositories of corresponding resources (digital items, digital essences, which are the effective files). Service Providers use the OAI-PMH interfaces of the Service Providers to collect and store their metadata as shown in [Xiaoming, 2001] and [Park, 2009]. They use the collected metadata for the purpose of providing one or more services across all the collected metadata like Pleiadi, Citeseer. The types of services which may be offered include a query/search interface, peer review system, cross linking, etc. Recently, an architectural shift was to move away from only human supporting end user interfaces for each repository, in favour of both human end-user interfaces and machine interfaces for data collecting.
It is well know that the resource discovery is the first step of the knowledge building. As explained in [Bellini, Nesi, 2009], at the moment it is very difficult for the user to know if a resource exists and it is available online, etc. In the OA domain, the access (in terms of discover and obtain resources) is still an open issue. In fact there are several causes that determinate the difficulties to disseminate the Open Access research prodices. The principals are three: a) the low quality of user interface design, the copyright

[3] DSpace <http://www.dspace.org>
[4] Fedora <http://www.fedora.info>
[5] EPrints for Digital Repositories <http://www.eprints.org>
[6] Greenstone University of Waikato <http://www.greenstone.org>

problems, and 3) the metadata quality. Scope of this work is to face the third factor identified by supporting institution to obtain an high level of metadata quality for their IR, trough an online quality assessment.
Thanks to the low effort required for the assessment (automatic) and the scalability of the technology infrastructure adopted, this service is particularly oriented to the institutions with high rate of content submissions (in particular thought self-archiving) and very low resources to manage and review the related metadata.
Therefore, the main goal of this work is to set up a Metadata Quality Certification service to support universities and research institutions to obtain and maintain an appropriate quality level of their IR in a very simples and economical way, defining:

a) A Metadata Quality Profile and related dimension able to be assessed through automatic processes.
b) A set of suitable metrics to be used as statistical tool for assessing and monitoring the IR implementation in terms metadata quality, trustworthiness and standard compliance.
c) A set of measurement tools to asses the metrics defined

Moreover, a MQC service is designed.

d) To achieve a better comprehension on Open Access implementation weakness, stimulating and directing new efforts towards technology, policy, standardization levels since the usage of the current widespread solutions is too vague to be exploited at a reasonable cost in the open world.

The possible benefits of a MQC service are:

1) A dissemination and exploitation growth of the IR research products.
This objective is related to the increase of the retrievability and accessibility of the research production deposited in a IR. This opportunity to freely access high quality research results can allow technology transfer between research institutions and industries.

2) The use of the IR research products in the national research evaluation process (CIVR)

The use of a transparent quality certification service for IR allows a more easy adoption of Open Access scientific production in the national research evaluation as auspicated by Open Access Group of CRUI. This opportunity is able to give a more exhaustive view of the amount and quality of institution research production.

3) A growth of institutions or researchers visibility.
The quality of the research production is the base of the credibility of a research institution such as the university and researchers involved. The possibility to have indexed on line the open access resource set up the possibilities to citations
To provide an effective irretrievability ad access to resources thanks to the high quality of metadata associated set up the condition of a

4) A cost decrement for maintaining an high level metadata quality. This objective aims to address the problem of the cost namely: the presence of appropriate expertise in the institution, availability of months- man, complexity of the assessment process, lack of defined roles and responsibilities in managing IR, etc. In this sense and automatic service of assessment can tackle these issues that represent some of the main risks of fault in managing an IR.

5) Increased awareness of bibliographic/citation standards by authors. Increased submission of publications with bibliographical references reflecting the accepted standards [Blake, Knudson, 2002]

Methodology

The research has taken the following steps:

a) Open Archive metadata quality issues analysis

The analysis of the metadata quality issues in the Open Archives repository was conducted gathering information through desk research, experiment results and the author experience in the field. In particular, desk research has dealt with articles and project

reports, the experiment results are mainly based on [Bellini, Deussom, Nesi,2010], and the author experience come from his participation in several related projects and working groups.

b) Metadata Quality Requirements

The quality requirements step concern the identification of the key functionalities that metadata have to support and are related to the scope of the Open Access repository. Then an overview of the state of the art of the quality frameworks is provided. The analysis starts form the Software Quality and Metadata quality concepts review. In particular the are been taken in to account the ISO 9124 [ISO/IEC 9124, 2001], [ISO 25000], NISO report [NISO, 2001] and several Metadata quality model as [Moen, et al, 1998], [Stvilia, Grasser, Twidale, 2007] and [Bruce, Hilmann, 2004].

b)Metadata Quality Profile

According to the Service requirements indentified in the previous section, this step defines the quality framework and the baseline quality of reference for the service. This section takes into account the CRUI guidelines, the FRBR metadata requirements and the survey results to define the baseline quality of reference. The baseline of quality define the weights to be associate to each field.

d) High Level Metrics and Low Level Metrics definition

A set of High Level metrics (quality dimensions) is defined according to QP. Then these HLM are translated into suitable LLM to be computed. This task follows the GQM approach and take as input the conceptual quality model defined in the Metadata Quality Framework defined in the Planning section.

e)Measurement plan

The measurement methods definition takes into account the ISO 15939 [ISO/IEC 15939, 2002] workflow and defines which criteria are adopted to calculate the metrics. In this section is defined a Measurement plan with the definition of base and the derived measurements and the tools used for measuring.

b) Metadata Quality Certification service definition

This section is devoted to design an online certification service defining scenarios through the Scenario Based Design techniques [Carroll, 1995], to extract user requirements and envisage new

functionalities. Moreover an overview of the most important initiatives on repository certification is provided. Once the service requirements are defined, it is necessary to define the entire service workflow. This research use the GQM approach [van Solingen, Berghout] [Basili, 2005] [Berander, Jönsson, 2006] to plan the workflow and service implementation.

g) Prototype implementation

This part describe the software prototype. In particular is provided deceptions of the system architecture, the assessment workflow, the database and the grid based rules developed. Moreover is provide and overview of the AXMEDIS GRID infrastructure and the third-party tools used to perform the measurement such as Pear Language Detect, Jhove and Aspell.

First Section

Metadata Quality issues and requirements

Chapter 1
Metadata Quality issues

1.1 Introduction

The Open Access movement is growing up among universities and research institutions. This initiative is based on two main declarations: the Budapest Open Access Initiative (BOAI) -2001 and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities -2003 signed by over 500 and 300 organizations respectively. As stated before, the objective is to push the organizations such as universities a research centre to make freely accessible the products of researches, in particular those funded by public funds. In order to make these intentions effective, the OA publication has set up two main approaches:

a) The Gold OA Publishing modality is referred to the possibility for the authors to publish in an open access journal that provides immediate OA to all of its articles on the publisher's website. In the latter case the metadata quality has to support the journal business thus the metadata produced are accurate. The metadata quality issues come from the former case where the metadata production is mainly delegated to the users;

b) The Green OA Self Archiving[Harnad, 2007] [Harnad, et al, 2004] where authors can publish in any journal and then self-archive a version of the article for free public use in their institutional repository [NISO, 2001] or domain repository such as arXiv. OA repositories do not perform peer review themselves. However, they generally host articles peer-reviewed elsewhere. OA repositories can contain also preprints7, post-prints8, or both.

[7] A preprint is any version prior to peer review and publication, usually the version submitted to a journal.
[8] A post-print is any version approved by peer review. Sometimes it's important to distinguish two kinds of post-print: (a) those that have been peer-reviewed but not

As described in [Brody, et al, 2007], about 15% of researchers – across disciplines and around the world – make their published research articles OA by 'self-archiving' (Green) them on the web of their own accord. In the UK, however, 5 out of the 7 public research councils (RCUK) (and several further private ones) now officially require their grant recipients to self archive their findings as a condition of funding ; and some UK universities are likewise beginning to require it.
In the Digital Agenda for Europe – Driving ICT innovation by exploiting the single market (Chapter 2.5.2.)– refers to effectively managed knowledge transfer activities and states that publicly funded research should be widely disseminated through Open Access publication of scientific data and papers.[9] According to this line, the European Commission is conducting a pilot[10] on Open Access to peer reviewed research articles produced in the context of the Seventh Research Framework Programme (FP7) to ensure the outcomes of EC funded projects are disseminated as widely and effectively as possible. The main aim is to guarantee maximum exploitation and impact in the world of researchers and beyond.

The importance of the Open Access movement is confirmed also by the results of a survey conduced by the European Association for Cancer Research (EACR)[11] that found that 59% of researchers say their work is often hindered by a lack of free access to research findings.
Moreover the survey found out also that: a) Internet is used by 94% of cancer researchers for professional activities every day, with the majority accessing PubMed and online journals daily or 2-3 times a week; b) nearly three quarters of survey respondents have published work in open access journals, indicating a growing acceptance of OA as a route to publication; c) the 88% of respondents said publicly-funded research should be available to be read and used without access barriers [Kenney, Warden]. Indeed

copy-edited and (b) those that have been both peer-reviewed and copy-edited. Some journals give authors permission to deposit the first but the not the second kind in an OA repository.
[9] http://www.openaire.eu/it/open-access/mandates-a-policies
[10] http://www.openaire.eu/en/component/attachments/download/4.html
[11] http://www.eacr.org/about/20110820_Open%20Access%20Future.pdf

the OA it is not a European initiative only. The Princeton University, for instance, has included in its statue the OA policy[12]:

"*At a September 19 meeting, Princeton's Faculty Advisory Committee on Policy adopted a new open access policy that gives the university the "nonexclusive right to make available copies of scholarly articles written by its faculty, unless a professor specifically requests a waiver for particular articles*."

Similarly, the ERC Scientific Council Guidelines for OA pushes the institution in adopting common policies and standards seeing that "*over 400 research repositories are run by European research institutions and several fields of scientific research have their own international discipline-specific repositories*[13]" such as PubMed Central, arXiv, DDBJ/EMBL/GenBank and RSCB-PDB/MSD-EBI/PDBj protein structure database.

In Italy, several universities are including in their statutes a clear reference to the OA and are promoting the implementation of Institutional Repositories (IR) to deposit the products of researches.

In the OA perspective, the authors even though they are subject to copyright can deposit copies of their finished articles in the archives and published them on any magazine at the same time. Moreover, there are evidences that this practice does not affect subscriptions to magazines.

If the subject-discipline circulates not refereed pre-prints or working papers in advance of publication (like Physics, or Economics), then these can be deposited. If an accepted method of communication is through conference papers (like Computer Science), then these can be deposited: similarly for fields that use book chapters or reports[14]. Other fields like Biomedicine only circulate refereed post-prints. Indeed, it is require that the IR has to tag the peer-reviewed material to make this status clear. The important point is that repositories reflect and support the existing research culture of the discipline.

The system works by these electronic versions of article, or eprints, being deposited into a database, or repository. These repositories are mainly administered by research institutions, which confer the advantage of allowing local support of users. Such institutional repositories share records about their content with service

[12] http://theconversation.edu.au/princeton-goes-open-access-to-stop-staff-handing-all-copyright-to-journals-unless-waiver-granted-3596
[13] http://www.openaire.eu/en/component/attachments/download/3
[14] http://www.driver-support.eu/oa

providers, who then offer search services to users across every record that they hold. This means that a researcher using a search service is searching across all repositories, not just individual ones. Once the researcher finds a record, then they can view the full-text direct from the IR. Examples of these services, built up on he OAI-PMH protocols, are CiteSeer[15] (information science), RePec[16] (research paper in economics), Pleiadi[17] (OAI resources), etc. One of the most important Service Providers is OAIster[18]. OAIster is a union catalogue of digital resources. It provides access to digital resources by "collecting" their descriptive metadata (records) using OAI-PMH on thousands of contributors. The proposal has tried to eliminate the so called 'dead ends' (collected records which do not link to an accessible digital resource) of the query results provided by OAI service providers. In fact users retrieve not only descriptions about resources, but they have access to real digital resources thorough the URL of the access pages of CMS (i.e., http://aei.pitt.edu/7400/).

As well as services which just search repositories, the full-text is also searched by Google, Yahoo and others[19]. There is no charge for using IRs. The process of deposition typically takes few minutes and consists of filling in a web-based form with metadata about the article (Green road); then attaching a pdf copy (or similar), and then submitting it to the repository administrator. IRs have help-systems and guidance: some institutions may offer personal assistance for the first few times you deposit. The process is quick and simple and will mean that the article is then available world-wide to a vastly increased readership.

Unfortunately, the enthusiasm for this initiative has accelerated the implementation of those repositories, neglecting the adoption of common policies, guidelines and standards. A survey of existing Open Access regulations, for instance, initiated among the **European Heads of Research Councils (EUROHORCs)** member organizations (MOs) in December 2007, demonstrated the great variety of Open Access policies among the EUROHORCs MOs and two thirds of them have introduced Open Access policies. In April

[15] Citeseer <http://citeseer.ist.psu.edu>
[16] RePec <http://repec.org>
[17] Pleiadi <www.openarchives.it/pleiadi>
[18] OAIster Home <http://oaister.umdl.umich.edu/o/oaister/>
[19] DRIVER EU- project <http://www.driver-support.eu/oa>

2008, the General Assembly of EUROHORCs agreed to recommend a minimal standard regarding Open Access to its Member Organisations [EUROHORC,2008]. Several studies conducted over the use of the metadata schema in Open Access repositories, confirm this fragmentary landscape in terms of the interpretation of these schema itself, the policy adopted, the frequency of use of a certain field, and so forth. Moreover, the Green road approach brings some issues related to the filling modality of the field because many authors are not expert on cataloguing or don't care about the information they are providing. Again, some OAI service provider provides advanced search functionalities to the users, but, since the physical access to the resource is not provided in the same request action by the IR systems, there is no guarantee of successful access because the record is no updated on the data provider and the objects cold be no longer available (broken link). These issues and many others are clearly matter of metadata quality.
In any case, the research findings on the OAI-PMH assessment for Open Access repositories [Bellini, Deussom, Nesi,2010] has outlined also that the IR quality implementation is a fragmented landscape that range form high level of quality implementation in terms of number of deposit and quality of metadata to very low quality initiatives that are still at experimental level.
In the middle there are a number of IR that are slow to grow in number of deposits because of it is not always clear what are OA rules, modalities of use, opportunities, objectives, etc.
For instance the final user (research that can deposit their products and students that can perform research on the IR) can have doubts about what type of content an IR can accept, what are the responsibilities, the IPR assigned, which impact the IR assure, etc.
Thus an IR can hold digital duplicates of published articles and make them freely available.

1.2 Metadata Quality issues in OA repositories

Currently, web crawlers index most web pages for search engines but only index an estimated 16% of the vast numbers of text and non-text digital objects available [Lawrence, 1999]. One method information providers use to solve the information indexing and retrieval problem is to create data about the digital objects and to

make that data searchable. The set of descriptions about the resource itself is called metadata. "Metadata is structured data about data that supports the discovery, use, authentication, and administration of information objects" [Greenberg, 2001][Hope]. Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information [NISO, 2001]. Presently there are a number of different types of metadata commonly classified with the 5 categories presented by NISO [NISO, 2001]: Descriptive metadata, Administrative metadata, Structural metadata, Right management metadata, and Preservation metadata.

1) Descriptive Metadata

The descriptive metadata describe and characterize the resources for the purpose of retrieval and identification. Examples are: title, abstract, authors, keywords, persistent identifier, etc. Presently, there are some guidelines available and developed during the EU projects such as DRIVER [20] and OpenAIRE [21] that aim to stimulate and support institutions in IR implementation. Other IR guidelines are provided by Italian Cultural Heritage Ministry for the legal deposit service provided by National Library of Florence[22]. A Working Group on metadata for IR promoted by CRUI is working also on metadata definition in order to determinate which are the fields required to describe a resource, which are mandatory or optional according to the resource types (e.g. pre-print, book, article, proceeding, etc.) .

2) Administrative metadata

The Administrative metadata provide information about access restriction, when and how the resource was created, the file format description, the origin of the content, the provenance, and other technical and administrative information. These metadata are useful for resources management and for guaranteeing their credibility through certification and tracking tools during their lifecycle and generally implemented in the DLMS (Digital Library Management System). An example of

[20] DRIVER EU-project <www.**driver**-repository.eu>
[21] OPENAIRE EU-project <www.**openaire**.eu/it>
[22] MIBAC - Deposito Legale <www.depositolegale.it>

this set e of metadata is the MAG schema[23] developed by the Italian Ministry of Cultural Heritage.

3) Structural Metadata

The structural metadata describe the composition and the relation of compound resources. Examples are: the definition of the page order within a chapter and can be referred to a digital collection or a single complex object such as METS, SCORM, MPEG21, etc. In the Best practice of structural metadata are identified six levels[24]:

1. No Structural Metadata: cultural heritage materials (single images)
2. Structural Metadata Embedded in a PDF Document: course reserves
3. Structural Metadata Defining File Sequence: books, journals, cultural materials.
4. Structural Metadata Defining Logical Components: books (journal example not provided, but also appropriate for journals)
5. Physical and Logical Structural Metadata Encoded in a Finding Aid: manuscript collections with digital files
6. Structural Metadata with Analyzed Page Layout: newspapers

The possibility to obtain a more expressive description of the resource using complex object format via OAI-PMH was explored in [Van de Sompel, Nelson, Lagoze, 2004]. Even if these information are useful for the OAI Service provider, for implementing advanced service of the metadata harvested from OAI Data provider, these information often are not available through OAI-PMH protocol as shown in [Bellini, Deussom, Nesi,2010] where the percentage of Open Archive that adopts a complex object format such as METS or MPEG21-DIDL are around 15% and 10% respectively. In fact, the possibility to obtain a richer representation of the resource is

[23] Metadati Amministrativi e Gestionali (MAG) http://www.iccu.sbn.it/upload/documenti/manuale.html

[24] Best practices for structural metadata http://www.library.yale.edu/dpip/bestpractices/BestPracticesForStructuralMetadata.pdf

demanded mainly to the data provider repository software setting (crosswalks).

4) Digital Right Management Metadata

The DRMM inform about which are the rights (exploitation, reuse, dissemination, etc) defined for that resource. These information can be processed by automatic tools that can delivery the entire object or a its part according to the DRM definition or can be a simple text description. Metadata schema such as Dublin Core, MODS, MARC21 and METS (METSrights) foresee a specific elements or sub-set of the schema to describe rights. For instance the 'Right' element of Dublin Core allows a simple text description of the rights related to the resource or a reference to an external service able to provide those information. Rights expressions can be more complex and communicate if the access to a content, that can be wrapped in secure containers, is permitted and under what conditions. Complex expressions could be based on MPEG Rights Expression Language (REL)25, XrML[26], Digital Property Rights Language (DPRL) 27or Open Digital Rights Language (ODRL)[28]promoted by Open Access community.

Other initiatives are CopyrightMD[29] an initiatives of California Digital Libraries that identifies an XML schema with a minimum set of elements able to identify the state of the right of a resource.

5) Preservation metadata

Preservation metadata, mainly driven by PREMIS [30]a standard promoted by Library of Congress, is the information that supports the processes associated with digital preservation. More specifically, it is the information necessary to maintain the viability, renderability, and understandability of digital resources over the long-term. Viability requires that the archived digital object's bit stream is intact and readable from

[25] http://www.xrml.org/reference/MPEG21_REL_whitepaper_Rightscom.pdf
[26] XrML - The Digital Rights Language for Trusted Content and Services - http://www.xrml.org/about.asp
[27] http://xml.coverpages.org/dprl.html
[28] http://odrl.net/
[29] California Digital Library, CopyrightMD. http://www.cdlib.org/inside/projects/rights/schema
[30] http://www.loc.gov/standards/premis/

> the digital media upon which it is stored. Renderability refers to the translation of the bit stream into a form that can be viewed by human users, or processed by computers. Understandability involves providing enough information such that the rendered content can be interpreted and understood by its intended users. Preservation metadata can serve as input to preservation processes, and also record the output of these same processes[OCLC/RLG, 2002].

In the self- archiving publication process, the authors have to provide only the descriptive metadata to catalogue their resources. The others have to be under institution control or are self generated by the IR system. Thus, this research focuses the descriptive metadata that are those used by the users to retrieve and access the digital objects and they can be harvested freely with OAI-PMH protocol.

OA fragmented landscape

In [Bu, Park] is described the case study of the assessment of Metadata Quality: on National Science Digital Library Metadata Repository. The metadata records generate in this repository, are used in the search engine (Search and Discovery by UMASS) to return results for a search. When the entire text of a resource cannot be accessed freely due to licensing issues, the metadata is likely the main source of information about this resource. Since incoming records do not go through a standardization process, the metadata submitted by the different organizations can vary greatly in quality. The results of an extended assessment performed on all OA registered on www.openarchive.org is presented in [Bellini, Deussom, Nesi, 2010] and confirm the presence of criticisms on metadata quality. In fact some criteria such as "Use of Metadata set" Show that the distribution of metadata set is quite spread.
Moreover, there exist a 15% of institutions using metadata sets which are personal model (single instances in the distribution) or which do not have a significant number of institutions. The adoption of non-standard metadata set and schema affects the effectiveness of archive visibility and distribution. Examples are URI schema: http://libst1.nul.nagoya-u.ac.jp/akf/akf.xsd or URIschema: http://uhasselt.be/agris/1.0.xsd.

The research findings of [Bellini, Deussom, Nesi,2010] provide a further confirm of the level of fragmentation of OA landscape analysing the present IR metadata implementations. There are several metadata standards promoted by different communities or by a single community itself to describe resources managed by OA. The most common are: Dublin Core [31](generally supported by default), METS[32], MPEG21 DIDL[33] (as a wrapper of other metadata models), MARCXML[34], etc. The table here below shows the first results of the harvesting of metadata sets from open archives around the world listed in the www.operarchive.org.

N	Prefix	Schema
100%	OAI_DC	http://www.openarchives.org/OAI/2.0/oai_dc.xsd
15%	MARCXML	http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd
15%	METS	http://www.loc.gov/standards/mets/mets.xsd
14%	Rfc1807	http://www.openarchives.org/OAI/1.1/rfc1807.xsd
14%	Oaimarc	http://www.openarchives.org/OAI/1.1/oai_marc.xsd
11%	MPEG21-DIDL	http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd
8%	RDF	http://www.openarchives.org/OAI/2.0/rdf.xsd
6%	Uketd_dc	http://naca.central.cranfield.ac.uk/ethos-oai/2.0/uketd_dc.xsd
6%	Junii2	http://ju.nii.ac.jp/oai/junii2.xsd
5%	Context_ob	http://www.openurl.info/registry/docs/info:ofi/fmt:xml:xsd:ctx
4,5%	Oai_etdms	http://www.ndltd.org/standards/metadata/etdms/1.0/etdms.xsd
4,5%	Xepicur	http://www.persistent-identifier.de/xepicur/version1.0/xepicur.xsd
4%	junii	http://ju.nii.ac.jp/oai/junii.xsd
3%	qdc	http://epubs.cclrc.ac.uk/xsd/qdc.xsd
3%	xmetadiss	http://www.d-nb.de/standards/xmetadiss/xmetadiss.xsd

It is also well known that metadata sets may be different for different domains, cultural background, scope, type of digital contents or business model. For instance, metadata sets required to catalogue physics resources can be different with respect to those used for media or ICT works, and again different from those adopted for administrative institutional documents, etc.
This lack of uniformity has generated several different standards and again for each of them, several different implementations

[31] http://www.openarchives.org/OAI/2.0/oai_dc.xsd
[32] http://www.loc.gov/standards/mets/mets.xsd
[33] http://standards.iso.org/ittf/PubliclyAvailableStandards/MPEG-21_schema_files/did/didl.xsd
[34] http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd

and/or personalization of metadata sets. In our analysis, 153 different metadata schemas have been identified, over only 853 repositories; thus a high percentage [Bellini, Deussom, Nesi,2010]. This count aggregates the records on the metadata Schema field (that is mandatory) from metadata format table. In Table 1, the percentages of the metadata sets' spread among the observed OAs are reported together with their schema and typical prefix. This stable reports the spread percentages of 16 most used metadata, with respect to the total number of different sets of 153. Noteworthy is that the DC is largely the most common, while after DC a number of metadata sets is in the range of 8-15% such as RDF, METS, MPEG-21, etc.

Issues in the schema implementations

When searching and browsing across archives, a user will expect those search capacities that are also provided in a single archive environment. He 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 between a certain period of time, 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 comparable, there must be a uniform date format and an ordering on that format.

Then, when the user is looking for an author, he is not interested in other information regarding the authors, thus, if the author field contains more than the name (address, affiliation), then the system has to distinguish between the author name and the rest of the author information[Fischer, Fuhr].

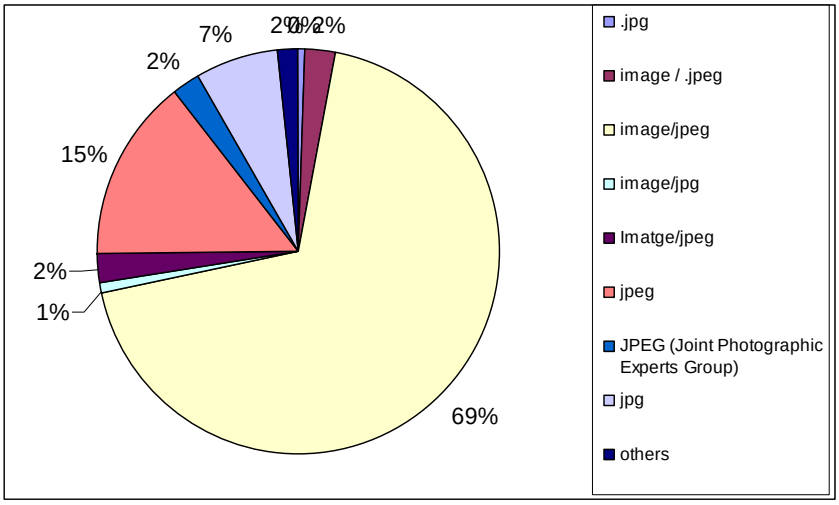
Metadata from the same single archive can be expected to have a uniform format for e. g. the author information, but in a domain with a low level of standardization, could be very difficult.

In [Bellini, Deussom, Nesi,2010] have been detected some problems regarding the interpretation and use of the single metadata fields. Moreover, it is well know that the use of simple Dublin Core foresees a high level of flexibility for filling in the metadata field. The performed analysis has shown that a very few number of institutions did adopt a qualified DC model, as defined by standard

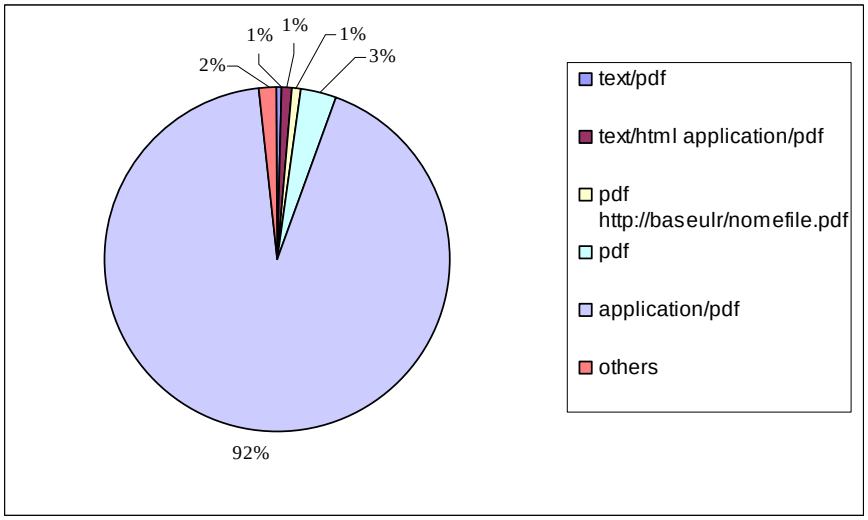
recommended best practice with a controlled vocabulary such as RFC 4646 or ISO639-1.
Moreover, the metadata multi-language system is managed in two modalities: using different instances of DC: language for each language or expressing different languages in the same field with a separator. The analysis has outlined that this separator can be arbitrary the sequent types: ` ,' ` ;' ` – ` '/'. Here below is provided an overview of the different instance of language founded in metadata harvested.

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

Regarding the DC:format field we have found different filling modalities with the presence of the file format definitions, physical medium descriptions, the dimensions of the resource and as described by standard definition. The recommended best practice refers to use a controlled vocabulary such as the list of Internet Media Types [MIME]. Here below we provide an example of the use of this field for JPEG and PDF file format.
Usage for JPEG file format: the right form is image/jpeg or image/jpg.



Usage of PDF file format: The right form is application/pdf



In fact, when looking for documents written, for example, in English, the user will not want to bother with guessing the different keywords for "English" ("eng", "English", "en us" etc.), he will just want to specify English as the document's language and leave the rest to the search service.

But if this is a matter of user interface, the search engine in background should not address all possible ways in which the vale can be written.

At the moment, the principal metadata standard adopted in IR, not include authority control system. For instance the OAI recommend the use of DC as a basic standard to implement the OAI-PMH transfer protocol, but DC not allows the distinction among control or

not controls form of the author name. In fact, the IRs born without the authority control tools on metadata of resources deposited and according to is stated in [Salo, 2009] the a) Self-archive resource and related self insert of metadata by the authors with this type of interface and b) the missing of automatic procedures for inserting the these data rise up the risk of collecting poor quality metadata in the IR.
Through examining learning objects and e-prints in [Barton, Currier, Hey, 2003] and [Guy and Powell, 2004], the importance of quality assurance for metadata creation is shown while pointing out the lack of formal investigation of the metadata creation processes such as inaccurate data entry (e.g., spelling, abbreviations, format of date [date of creation or date of publication], consistency of subject vocabularies) that result in adverse effects on resource discovery.
In Open access domain, the metadata quality affects not only the service offered through the archive's native Web interface, but also what options can be offered by OAI service providers like OAIster, Pleiadi, etc. The usefulness of a digital repository is strongly correlated to the quality of the metadata that describe its resources.

According to [Kelly, Closier, Hiom] and the findings of the analysis reported here, the main difficulties of the Open Access domain to be complain to standard and guidelines and consequently to be effective in disseminating the research produces can be summarized as follow:

- in some cases the 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.
- Concerns over changes in standards during the projects' lifetime.
- Software tools and interfaces not suitable
- Not well defined mandate (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 an IR.

It is clear that Metadata quality is an open issue for open access community and in order to tackle these weaknesses a metadata quality certification service is needed.
In the next chapter are identified the main quality requirements necessary to define a quality profile for the Open Access metadata.

Chapter 2

Metadata Quality Requirements

2.1 Introduction

Presently, different domains tackle the quality issues from the process or product point of view. In [Garvin, 1984], David Garvin studied how quality is perceived in various domains, including philosophy, economics, marketing, and operations management. The results were that "quality is a complex and multifaceted concept" that can be described from five different perspectives:

a) The *transcendental view* sees quality as something that can be recognized but not defined.

b) The *user view* sees quality as “fitness for purpose”. This view of quality evaluates the product in a task context and can thus be a highly personalized view. In reliability and performance modelling, the user view is inherent, since both methods assess product behaviour with respect to operational profiles (that is, to expected functionality and usage patterns) [Kitchenham, Pfleeger, 1996].
In view is adopted in [Guy and Powell, 2004] where in the context of metadata “fitness for purpose” means that the high quality of metadata are able to support the functional requirements of the system is designed to support.

c) The *manufacturing view* sees quality as conformance to specification. This view examines whether or not the product was constructed "right the first time," in an effort to avoid the costs associated with rework during development and after delivery. This process focus can lead to quality assessment that is virtually independent of the product itself. That is, the manufacturing approach adopted by ISO 9001 and the Capability Maturity Model advocates conformance to process rather than to specification [Kitchenham, Pfleeger, 1996]. Manufacturing production is a process that takes place in a controlled environment (factory). Instead, the nature of the OA publication process is different because the content

submission of the to an IR could be done out of a controlled environment (OA green road) like the university library dept. Therefore, since the result of the submission process is mostly unpredictable, the *manufacturing view* is not applicable in our case.

d) The *product view* sees quality as tied to inherent characteristics of the product. This approach is frequently adopted by software-metrics advocates, who assume that measuring and controlling internal product properties (internal quality indicators) will result in improved external product behaviour (quality in use). Assessing quality by measuring internal properties is one of the objectives of the present work. [Kitchenham, Pfleeger, 1996].

e) The *value-based view* sees quality as dependent on the amount a customer is willing to pay for it.

Following the product point of view, there are several standards that define the Quality concept in the software domain such as ISO/IEC 9126. An ISO/IEC 9126-1 quality model defines a set of characteristics or dimensions which are further refined into sub-characteristics which in turn are decomposed into attributes [Botella], [ISO/IEC 9124, 2001]. These main characteristics are: functionality, reliability, usability, efficiency, maintainability and portability. At the end of the hierarchy there are suitable metrics that might be designed adopting a particular approach like Goal Question Metric paradigm [Basili, Caldiera, Rombach, 1994]. The new ISO25000 SQuaRE package replaces the ISO/IEC 9126 series and ISO/IEC 14598 series providing a comprehensive view of the Quality including the Quality framework and the evaluation process (Software Products Quality Evaluation Reference Model) and describing activities and tasks to be performed during the quality evaluation of the products.

Following the Fitness for purpose point of view, the [Guy and Powell, 2004] work considered high quality metadata if support the functional requirements of the system it is designed to support. They defined internal and external functional requirements of metadata in relation to the archive's web user interface such as search, browse, filter by, etc. These functional requirements are used to decide what metadata are needed so that the metadata quality can be assessed defining whether the metadata in Eprint

archive are good enough to support these functional requirements according to the aim of the archive, the designed community[35], the type of objects you are going to manage, and so forth.
In [Evans, Lindsay], the quality definition in related to the meeting or exceeding customer expectations or satisfying the needs and preferences of its users put more emphasis on user needs.
As stated in [Margaritopoulos, 2008], the relevance of metadata of a resource, and consequently their quality, has to determinate in their **context of use.** For instance a metadata record of absolute correctness and full completeness may not be of quality if the values of metadata fields do not comply with the context of use (domain standards and guidelines).
The completeness itself can be assessed in different way because a metadata might be required in a certain domain and does not in another and furthermore different domains can define even different encoding for the same field.
In the Building quality assurance into metadata creation [Barton, Currier, Hey, 2003] is described that the metadata quality, the semantic and descriptive elements associated to each resource in a IR, affects the quality of the service provide to the IR users.
Similar to these approaches that identifies the metadata requirements in relation to the final user expectations, are the [ISO 14721, 2003] and [Park, 2009] approaches. In [Park, 2009], is described how the quality of metadata affects the bibliographic function of research, use, dissemination, authenticity and management. In fact the article defines that the main scopes of the metadata are related to retrieval, identification, selection and delivery of resources that are the main functions of online catalogues and digital libraries.
In the Open Archive Information System (OAIS) standard [ISO 14721, 2003], the Generate Descriptive Information function extracts Descriptive Information from the Archive Information Packages (AIPs) and collects Descriptive Information from other sources to provide to Coordinate Updates, and ultimately Data Management. This includes metadata to support searching and retrieving Archive Information Packages (AIPs) (e.g., who, what, when, where, why), and could also include special browse products (thumbnails, images).

[35] ISO OAIS 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.

The NISO specifies that an important reason for creating descriptive metadata is to facilitate discovery of relevant information and can help organize electronic resources, facilitate interoperability and legacy resource integration, provide digital identification, and support archiving and preservation [NISO, 2001].

This research addresses the metadata quality in the **context of the Open Access IR** with the aim of supporting the institutions in improving and maintaining high level of metadata quality for their contents. In fact, a low level of metadata quality affects the possibility to discover and access these resources preventing their effective reuse and dissemination and losing the benefits of being Open Access content.

2.2 Metadata Quality definition

The metadata quality issue is still relatively unexplored, but there is growing awareness of the essential role of metadata quality to exploit contents in the repositories. In fact, the creation of metadata automatically or by authors who are not familiar with 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 it difficult to locate relevant information. While there is a wide agreement on the need to have high quality metadata, there are fewer consensuses on what high quality metadata means and much less in how it should be measured.
Quality is defined in the ISO 8402 - 1986 as: "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs".
This definition includes the user perspective (needs) and product characteristic perspective (features) but we aim to highlight the importance of the community served in defining the metadata quality in the OA context, adopting a different definition.
Thus, we assume the metadata quality definition as "fitness for purpose" because are fixed by the domain not only the stated purposes of the metadata but also their relevant features (metadata schema to be used, guidelines, etc.).

Hence in order to unambiguously evaluate the quality of these metadata against the domain objectives, it is necessary to break down the context purposes into specific functionalities with defined characteristics, and then, link these functionalities to respective quality dimensions and measurable metrics.

In literature, these functionalities, quality dimensions and metrics definitions are in general presented in a comprehensive Quality Framework (QF). Several researchers have addressed the information quality issues developing QFs. These QFs define several dimensions that information should comply in order to be considered of high quality. As already stated in [Ochoa, Duval], these QFs vary widely in their scope and goals. Some have been inspired by the Total Quality Management (TQM) paradigm, such as [Strong, Lee, Wang, 1997]; others are from the field of text document evaluation, especially of Web documents, such as [Zhu, Gauch, 2000] others are liked to degree of usefulness or "fitness for use" [Jura, 1992] 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 [NISO, 2007]:

1) Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current and likely use of the digital object.
2) Good metadata supports interoperability.
3) Good metadata uses standard controlled vocabularies to reflect the what, where, when and who of the content.
4) Good metadata includes a clear statement on the conditions and terms of use for the digital object.
5) Good metadata records are objects themselves and therefore should have the qualities of "archieability", persistence, unique identification, etc. Good metadata should be authoritative and verifiable.
6) Good metadata supports the long-term management of objects in collections.

These criteria and principles are defined by NISO to provide a framework of guidance for building robust digital collections but they do not provide a clear number of well defined quality dimensions leaving the implementers free to address the issues in different ways. For instance, "supporting of interoperability" and the "using of authority control and content standards", are requirements that

without a formal definition they can be only considered from the not computable and "transcendental" point of view.
There are other metadata QFs that are formally defined and can be computed. They differ in granularity/detail, name of dimension, complexity and operational but there are many overlaps among them. In [Stvilia,2006] are identified three types of approach to studying information quality: 1) intuitive, 2) theoretical, and 3) an empirical approach.
The intuitive approach is identified when the researcher selects information quality attributes and dimensions using his intuition and experience. In theoretical approach, quality dimensions 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 the user applies for assessing information quality.
In [Wang, Strong] is explained that the theoretical and intuitive approaches concentrate more on information products development quality rather then on quality in use. The ability of selecting the dimensions of quality most relevant to a particular study was identified as the advantage of using an intuitive method, while the potential of producing a comprehensive list of quality dimensions was named as the string side of the theoretical model. The empirical approach starts from a user survey asking to them to name the dimensional and attributes coming to mind when they think about quality [Stvilia,2006].

In [Moen, et al, 1998] are identified 23 quality parameters. However, some of these parameters (ease of use, ease of creation, protocols, etc) are more focused on the metadata schema standard or metadata generation tools. Given that the metrics should be technology-agnostic and measure only the quality of metadata instance, in this work we have followed a different approach for defining the Quality Framework.
Stvilia [Stvilia, Grasser, Twidale, 2007] uses most of Moen's parameters (excluding those not related with metadata quality), add several more, and group them in three dimensions of Information Quality (IQ): Intrinsic IQ, Relational/Contextual IQ and Reputational IQ. As defined in [Stvilia, et al] each dimension is described as follow:
1. Intrinsic IQ: is related to attributes that can be measured in relation to a reference standard. Examples include spelling mistakes

(dictionary), conformance to formatting or representation standards (HTML validation), and information currency (age with respect to a standard index date, e.g. "today"). In general, Intrinsic IQ attributes persist and depend little on context, hence can be measured more or less objectively.

2. Relational/Contextual IQ: This category of IQ dimensions ensures relationships between information and some aspects of its usage context. One common subclass in this category includes the representational quality dimensions – those that measure how well an information object reflects some external condition (e.g. actual accuracy of addresses in an address database). Since metadata objects are always surrogates for (hence bear a relationship to) other information objects, many relational dimensions apply in measuring metadata quality (e.g. whether an identifier such as URL or ISBN actually identifies the right document; whether a title field holds the actual title). Clearly, since related objects can change independently, relational/contextual dimensions of an information item are not persistent with the item itself.

3. Reputational IQ: This category of IQ dimensions measures the position of an information artefact in cultural or activity structure, often determined by its origin and its record of mediation.

The Stvilia et al. framework describes 32 parameters in total and some of the parameters (accuracy, naturalness, precision, etc) are present in more than one dimension Bruce & Hillman [Bruce, Hilmann, 2004], based on previous Information Quality research, condense many of the quality parameters in order to improve their applicability. They describe seven general characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. A relation between the frameworks of Bruce & Hillman and Stvilia et al. is proposed and summarized in Figure1 [Bruce, Hilmann, 2004].

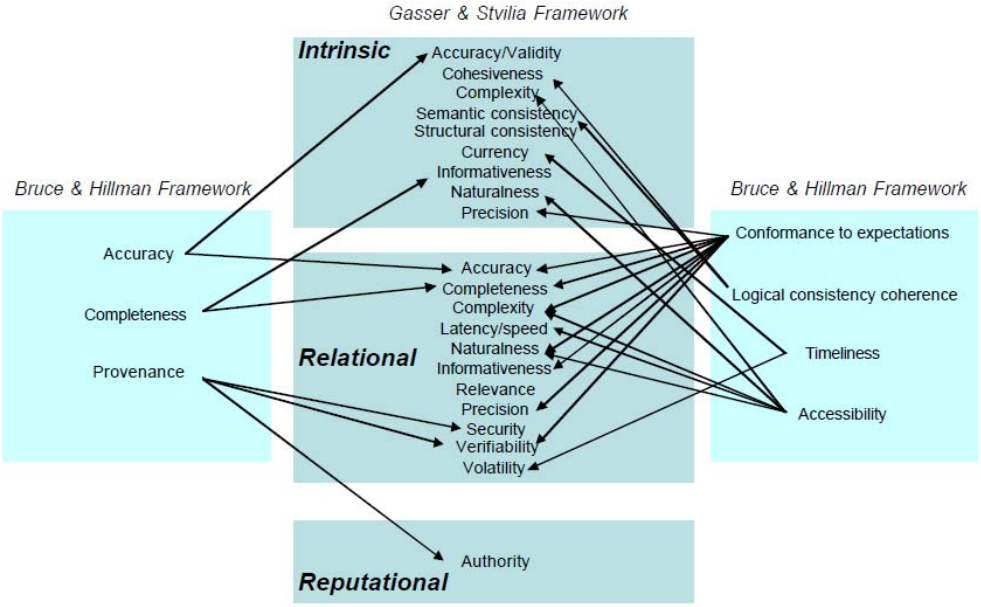


Figure 1 - The relation between Bruce & Hillman and Stvilia et frameworks.

However, these definitions are oriented toward the same directions; in fact user information needs are mostly driven by action/task the user requires to perform and that are well represented by FRBR requirements descried in the next paragraph. In fact, at first glance we can say that the quality of metadata reflects the degree to which the metadata perform the core bibliographic functions of find, identify, select and obtain a digital resource [IFLA, 1998]. In the next paragraph is defined the quality profile for the OA metadata and are accommodated all these considerations.

Second section

Metadata Quality Framework

Chapter 3
Metadata Quality Framework

3.1 The Goal-Question-Metric approach

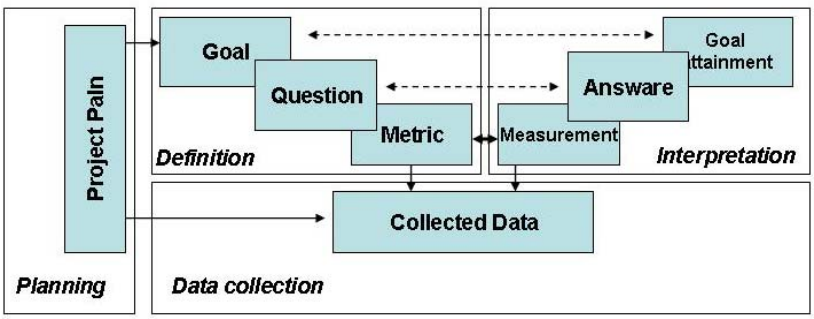
In order to address that transparency and objectively required for a quality assessment, the adoption of a standard methodology for design metrics and manage the entire workflow is crucial.
Although GQM originated as a measurement methodology for software development, the basic concepts of GQM can be used anywhere that effective metrics are needed to assess satisfaction of goals [Basili, Caldiera, Rombach, 1994]. The GQM paradigm represents a practical approach for bounding the measurement problem. It provides an organization with a great deal of flexibility, allowing it to focus its measurement program in its own particular needs and objectives. It is based upon two basic assumptions:
1) that a measurement program should not be 'metrics-based' but goal-based and, and
2) that a definition of goals and measurements need to be tailored to the individual organization.
This assumption requires that the organization (in our case the OA domain) makes explicit its own goals/purpose.
The literature [Basili, Caldiera, Rombach, 1994] typically describes GQM in terms of a six-step process where the first three steps are about using business goals to drive the identification of the right metrics and the last three steps are about gathering the measurement data and making effective use of the measurement results to drive decision making and improvements. In [Basili, 2005], are described the GQM six-step process as follows:

1. Develop a set of corporate, division and project business goals and associated measurement goals for productivity and quality.
2. Generate questions (based on models) that define those goals as completely as possible in a quantifiable way.

3. Specify the measures needed to be collected to answer those questions and track process and product conformance to the goals.
4. Develop mechanisms for data collection.
5. Collect, validate and analyze the data in real time to provide feedback to projects for corrective action.
6. Analyze the data in a post-mortem fashion to assess conformance to the goals and to make recommendations for future improvements.

Once appropriate metrics are identified, the last three steps of the GQM process address how to implement the metrics program in a way that ensures the focus will remain on goal attainment. In [Van Solingen, Berghout] these 6 steps are compressed in the following four phase that this work has used as a base of the entire research workflow:

1) **The planning phase**, during which a project for measurement application is selected, defined and planned in a project plan.
2) **The Definition phase**, during which the measurement programme is defined (goal, questions, metrics, and hypothesis are defined).
3) **The Data Collection phase**, during which actual data collection take place, resulting in a collected data.
4) **The Interpretation phase**, during which collected data is processed with respect to the defined metrics into measurement results, that provide answers to the defined metrics into measurement results, that provide answer to the defined questions, after which goal attainment can be evaluated.

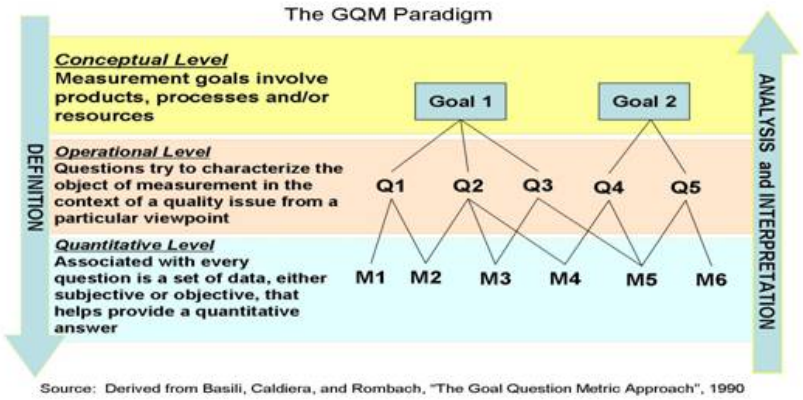


Planning Phase

In this work, this phase is represented by the MQP definitions. The GQM approach defines two types of goals: business goals and measurement goals where business goals drive the identification of measurement goals. It is not important whether the business goals are developed under the umbrella of GQM, or as a function of strategic planning. As we stated, the MQP is based on the goal or purpose of metadata records into the OA domain. The MQP drives the metrics definition. The MQP must exist because without them, the measurement program has no focus.

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). GQM defines a measurement model on three levels as illustrated in the figure below:



Basili expresses GQM goals (measurement goals) using five facets of information to define what the measurement should accomplish in precise terms. Each GQM goal statement explicitly contains these facets:

- Object: The process under study; in our case Completeness, Accuracy and Consistency.
- Purpose: Motivation behind the goal.
- Focus: The quality attribute of the object under study (what).
- Viewpoint: Perspective of the goal.
- Environment: Context or scope of the measurement. program.

The purpose of Basili's is to clarify and refine the measurement goals, moving from a conceptual level to an operational level by

posing questions. By answering the questions, one should be able to conclude whether a goal is reached. Once goals are refined into a list of questions, metrics are defined to provide all the quantitative information to answer the questions in a satisfactory way.
The term can mean a base measure, a derived measure, a composite or aggregate of measures, or, what some would call, an indicator.
At this point it is important to make a distinction between the metrics that are defined and the data elements that support them. The metric is at a more abstract level than the actual data items and measurements that need to be collected to provide the correct data for preparing the metric.

Data collection phase

Once the 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 ISO/IEC 15939 Measurement Information Model and includes:

- the definitions of direct measurements (ISO/IEC 15939 base measurement) with all possible outcomes (values),
- the medium (tools) that should be used for collecting the measurement (ISO/IEC 15939 measurement methods).
- the definition of derived measurement

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. Moreover, a quality improvement effort is estimated

3.2 Quality Profile definition

As we stated before, every quality assessment requires a definition of a clear and stable baseline quality of reference called **Quality Profile (QP)**.

A QP allows of taking into account the user perspective in the definition of the baseline quality of reference. For instance, in [Burnett, Ng, Park, 1999] a study on six metadata standards is presented and outlines that title, author, and identifier are common to all the schemes, and that two others – place and date – are common to five of the six schemes. This might imply that the impact of these fields in the overall metadata quality estimation is stronger than other fields but a formal definition is needed.

Thus, the first step is to remark the "purpose" of the Open Access IR:
"The purpose of Open Access (OA) is to maximise research access, usage and impact, thereby maximising research productivity and progress, in the interests of research, researchers, their research institutions, their research funders, the R&D industry, students, the developing world, and the tax-paying public for whose benefit research is funded and conducted."[36]

Hence, the QP has to be defined through the identification of the metadata functionalities in the Open Access IR domain and an evaluation of the user perspective. 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.

QP- FRBR based
In order to address this requirement has taken into account the FRBR [IFLA, 1998] model and the ICP International Cataloguing Principles promoted at IFLA 2009. In the final report are identified four main tasks performed by the users when searching and making use of national bibliographies and library catalogues:

- using the descriptive metadata **to find** materials that correspond to the user's stated search or discovery criteria (e.g., in the context of a search for all documents on a given subject, or a search for a recording issued under a particular title).

[36] On Sat, 31 Mar 2007, in response to "Mobilising Scholarly Society Membership Support for FRPAA and EC A1," Fred Spilhaus, Executive Director, American Geophysical Union, wrote, in the American Scientist Open Access Forum: http://openaccess.eprints.org/index.php?/categories/12-Learned-Societies

- using the descriptive metadata retrieved **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;
- using the descriptive metadata **to select** a resource that is appropriate to the user's needs (e.g., to select a text in a different language or version);
- using the descriptive metadata in order **to acquire or obtain** access to the resource described (e.g. to access in a reliable way to an online electronic document stored on a remote computer).

The results of this analysis show that the metadata functional requirements can be taken as baseline parameters to determinate the QP of IR metadata. Under this point of view a low level of metadata quality in a repository affects the capability of addressing the FRBR requirements defined above.

The IFLA Cataloguing section Working Group on the Use of Metadata Schema studied a "common core" of metadata elements to be recommended to libraries and catalogue agency [IFLA, 2003]. In fact one of the main objective of the Working Group was *"to determine a metadata" core record" – i.e., a set of most commonly occurring elements in selected metadata schemas – that could be used by authors and/or publishers of electronic records to enhance resource discovery, and to provide, where appropriate, elements for incorporation into bibliographic records (catalogue records).* The Working Group aimed to make recommendations as to which elements would be mandatory versus optional for both electronic serial and integrating resources and monographic resources. The analysis started from the eight areas of the International Standard Bibliographic Description (ISBD), and the fifteen elements of the Dublin Core metadata schema with the scope to find out a baseline set of constituent named metadata elements for describing any electronic resource in any domain regardless of the metadata schema used (i.e., schema-independent). The Working Group compiled a list of ten metadata elements that could be included in an FRBR-compliant record. Behind to each element we mapped the corresponded DC field according to the element description [IFLA, 2003]:

- Subject: what a resource is about (DC:description, DC:subject).
- Date: A date associated to the resource, e.g. creation date (DC:date).
- Conditions of use: Indicates the limitations and legal rules that may restrict or deny access to a resource (DC:righs).
- Publisher: Information about the entity responsible for making resource available (DC:publisher).
- Name assigned to the resource: The title of the resource (DC:title).
- Language/mode of expression: identify the language of the resource (DC:language).
- Resource identifier: Unique name or code to identify the resource (DC:identifier).
- Resource type: it includes both carrier and type of content (DC:type, DC Format).
- Author/creator: Name(s) of organization(s) or individual(s) responsible for creating or compiling the intellectual or artistic content of the work (DC:author, DC:contributor).
- Version: Provides information on the version, edition, or adaptation of a particular work, or relationships to other works (DC:relation).

The FRBR requirements has been translated into a weights definition. In particular, the table below, starts from the benchmark presented in the IFLA report [IFLA, 2003] and translate the functional requirements in a *coverage* index used for estimate the weight.

DC \ FRBR	Identify	Select	Find	Obtain	FRBR-W
contributor	x		x		0,529
coverage		x			0,188
creator	x	x	x	x	1,000
date	x	x		x	0,677
description	x	x	x		0,717
format	x	x		x	0,677
identifier	x		x	x	0,812
language		x	x		0,512
publisher	x	x		x	0,677
relation		x			0,188
rights		x		x	0,471

source	x				0,206
subject	x	x	x		0,717
title	x	x	x	x	1,000
type	x	x		x	0,677

The coverage index is estimated in this manner:

$w_{Idenify} = 1/11 = 0,09$

$w_{select} = 1/11 = 0,09$

$W_{find} = 1/6 = 0,17$

$w_{obtain} = 1/9 = 0,1$

$Sum = w_{Idenify} + w_{select} + W_{find} + w_{obtain} = 0,45 = 1$

FRBR-W= $w_{Idenify} + w_{select} + W_{find} + w_{obtain}$ / SUM(W)

QP - CRUI guidelines based

Another input comes from CRUI with "IR metadata guidelines" report just delivered in draft version. The report identifies a different status for the DC fields in relation to the type of document (e.g. article, monographs, and so forth). Moreover, since the IR the have to support the oai_dc prefix to disseminate the metadata through OAI-PMH v2 in not qualified DC format, it is necessary a "dumb down" process which results in a mapping of the DC qualified to the DC not qualified. Thus, if two or more fields are mapped into a unique DC not qualified field, this one takes the status from the field mapped with the highest level of importance. For instance, if the dc:title.alternative and dc:description.abstract are defined as Optional but the DC:title is defined as Mandatory (M), the not qualified DC:title results with a Mandatory (M) status.

.

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

Indeed, the translation from the CRUI guidelines into weights is an approximation useful only to allow a comparison among different QP and may have been possible errors in weight estimation.

QP – OA User community based

As already stated, the QP definition has to be defined not only though the identification of the metadata functionalities in the Open Access IR domain but also through an evaluation of the user perspective.

In order to address this requirement, we submitted a specific questionnaire to the OA community with the aim of gathering their points of view about relevance of each DC field in a DC record quality assessment. Since the OA community is mainly oriented towards Universities and research institutions, we have identified Researchers 20,6%, Professors 12,7%, ICT experts15,9%, Archivists15,9%, Librarians 25,4% and students 9,5%, as our target. The questionnaire results are reported in Annex I.

Data Filtering

In order to be more confidence in the analysis, we filtered out the answers with the following criteria:

Critical target

The OA publication involves Researchers professors ICT experts on the side of submission and Librarian and Archivists on the side of publication management, while the students are usually less concerned. In this scenario the probability that the answers collected from the Students can represent a "noise" in the statistics is high.

Low 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 (Red area highlighted in the figure

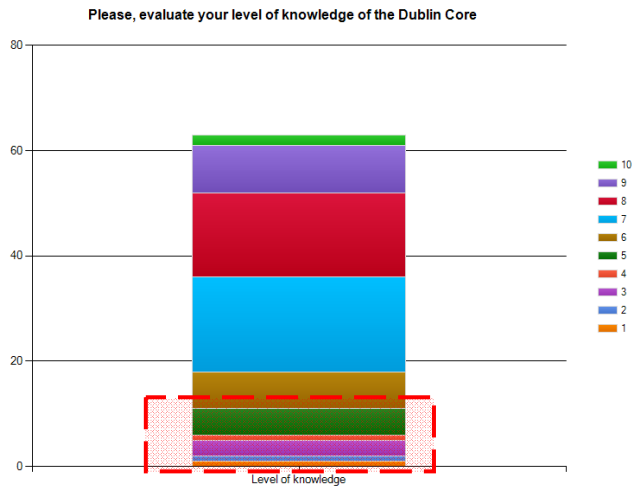


Figure 1 – Survey results on DC level of knowledge

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

Field selection

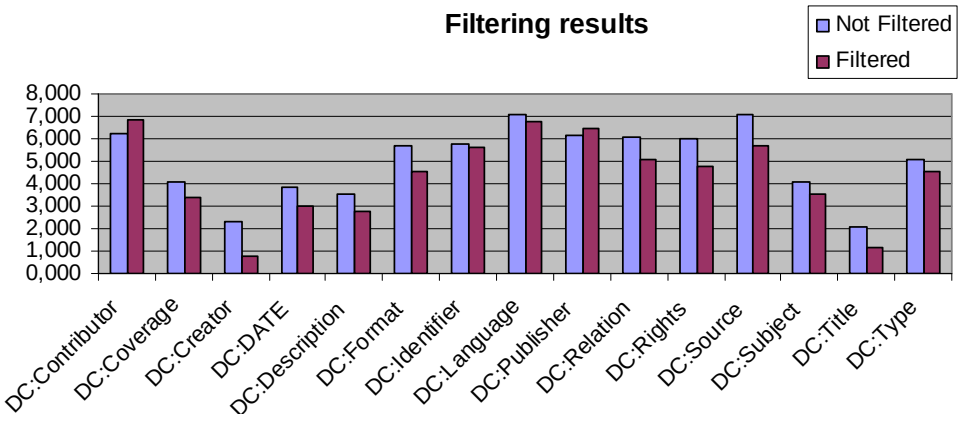
In order to define the quality profile we must determinate which are the fields to be assessed. In fact, each field has a different level of importance in a record. This level of importance has been estimated asking to the OA community to to assign a WEIGHT to each DC field from: 1 (the field can be omitted without affect the use of the record) to 10 (absolutely mandatory, the lack of the field makes the record totally unusable).

We assumed that the range from 1 to 5,5 can be considered as not important, thus we defined the following criteria to exclude those fields that are not considered determinant for the OA community, from the quality assessment:

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

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

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, reported in the graph, shown a reduction of the Variance for each field after the data filtering as a proof of the correctness of our assumption. Only in the case of Contributor and Publisher the Variance rises up.



In the table 1 are reported the Average (Avg), Media (Med), Standard Deviation (σ), Variance (σ^2) and Level of Confidence (Conf) for each DC field.

	Data not filtered					**Data Filtered**					
	Avg	Med	σ	σ^2	Conf (95%)	Avg	Med	σ	σ^2	Conf (95%)	Avg-Conf
Contributor	6,88	7,0	2,50	6,26	0,63	6,76	7,0	2,61	6,82	0,81	5,94
Coverage	5,66	6,0	2,01	4,06	0,50	5,90	5,5	1,83	3,35	0,57	5,33
Creator	9,30	10,0	1,53	2,34	0,38	9,50	10,0	0,89	0,79	0,27	9,22
DATE	8,57	9,0	1,95	3,82	0,49	8,61	9,0	1,72	2,97	0,53	8,08
Description	7,90	8,0	1,89	3,57	0,47	7,78	8,0	1,67	2,80	0,52	7,26
Format	6,87	7,0	2,37	5,66	0,59	6,61	7,0	2,12	4,53	0,66	5,95
Identifier	8,15	9,0	2,40	5,78	0,60	7,97	9,0	2,36	5,58	0,73	7,24
Language	6,71	7,0	2,654	7,04	0,66	6,59	7,0	2,60	6,78	0,81	5,78
Publisher	6,41	7,0	2,47	6,11	0,62	6,11	6,0	2,54	6,49	0,79	5,32
Relation	5,38	5,0	2,46	6,07	0,62	5,11	5,0	2,25	5,08	0,70	4,41
Rights	7,19	8,0	2,45	6,02	0,61	6,95	7,5	2,18	4,77	0,68	6,27
Source	6,01	6,0	2,65	7,04	0,66	5,66	5,5	2,38	5,69	0,74	4,92
Subject	7,47	8,0	2,02	4,09	0,50	7,33	7,5	1,88	3,54	0,58	6,74
Title	9,38	10,0	1,44	2,07	0,36	9,50	10,0	1,06	1,13	0,33	9,16
Type	7,14	8,0	2,24	5,06	0,56	7,23	8,0	2,13	4,57	0,66	6,57

Table 1 Data descriptive statistic results

According to the field selection criteria defined, the results show that Coverage, Publisher, Relation and Source have not passed the threshold of 5.5 (Figure 2). 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.

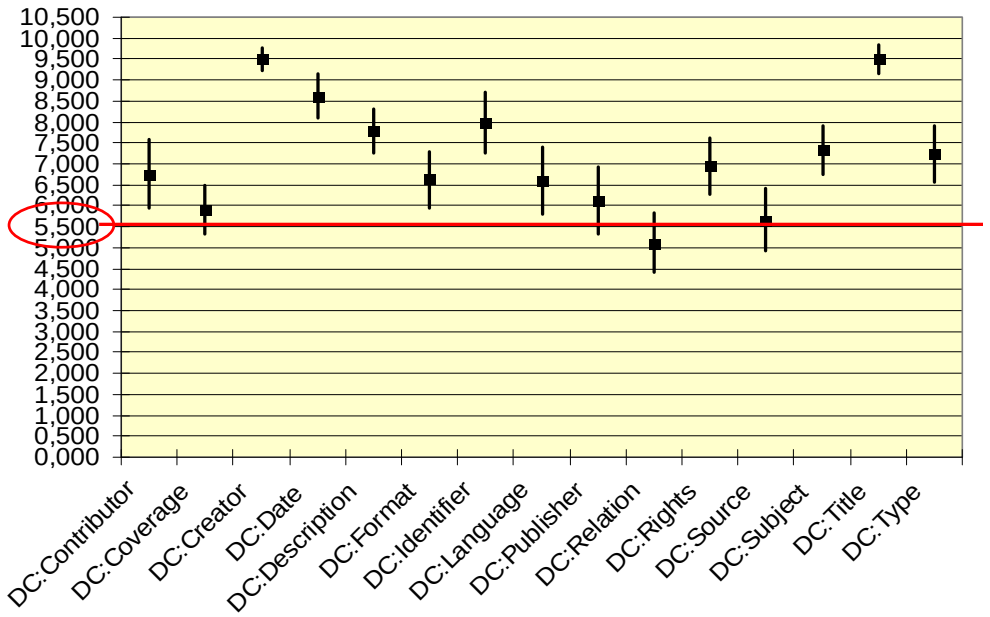


Figure 2 – Graph of the levels of confidence

This assessment allows a reliable selection of the fields to be evaluated through the quality assessment. The weights assigned to each field are the normalized Averages of the weights assigned by the users. The weights associated to the fields are reported in the table 2.

	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

Table 2 – MQC Quality profile

Now we have presented three possible QPs. In the table below is reported the values of correlation among the quality profile translated by CRUI guidelines, FRBR model and our MQC profile. The result show a good correlation between MQC and FRBR while between CRUI guidelines and the others there is not any correlation. This is due to some differences in evaluating in the importance of some DC fields for describing an OA resource.
Instead, the high correlation between the MQC and FRBR (green cell) allow us to say that MQC profile addresses also the FRBR metadata requirements.

	CRUI	*FRBR*	*MQC*
CRUI	1		
FRBR	0,479	1	
MQC	0,547	0,873	1

Correlation table

This table reports the distribution of the Averages of the DC fied usage assessed in several researches such as DISIT [Bellini, Deussom, Nesi,2010], Park [Bu, Park, 2006], Efron[Efron, 2007], Stvilia [Stvilia, et al. Obviously these results cannot be comparable since they are based on different population, hence, the aims is to gain a general overview on DC field usage and to explore

DC	DISIT	Park	Efron	Stvilia
contributor	0,18	0,08	0,35	0,06
creator	0,63	0,83	0,79	0,5
date	0,83	0,86	0,99	0,43
description	0,51	0,83	0,69	0,47
format	0,52	0,43	0,93	0,69
identifier	0,81	0,99	0,99	0,99
language	0,64	0,38	0,94	0,55
rights	0,29	0,16	0,18	0,41
subject	0,53	0,77	0,64	0,73
title	0,91	0,99	0,99	0,8
type	0,73	0,75	0,86	0,76

	DISIT	*Park*	*Efrom*	*Stvilia*	*CRUI*	*FRBR*	*MQC*
DISIT	1						
Park	0,847	1					

Efron	0,886	0,701	1				
Stvilia	0,714	0,676	0,641	1			
CRUI	0,794	0,594	0,652	0,428	1		
FRBR	0,635	0,801	0,532	0,496	0,479	1	
MQC	0,623	0,74	0,395	0,226	0,547	0,873	1

This table shows that the MQC have a low or neither correlation with the Researches results. This is positive since the DISIT results, for instance, comes out from an analysis of a random IRs where bad and best practices were included. Thus, if your dataset are from different sources, a low level of correlation or an correlation allows a real assessment. Instead, if the correlation is high with such datasets, it is possible that the QP might be based on less constraints. In other words, if a QP considers only few fields important, the probability to be more correlated to a random datasets grows.

3.2 High Level Metrics (HLM) definition

Completeness

There are different positions on the concept of Completeness defined in the Quality Frameworks analysed above. Commonly the Completeness is related to the empty field in a record set and is generically defined as the degree to which value are present in the attributes that require them [Pipriani, Ernst]. In [Bruce, Hilmann, 2004] 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. It is almost always possible to imagine describing things in more detail, but it is not always possible to afford the preparation and maintenance of more detailed information. Second, the element set should be applied to the target object population as completely as possible; it does little good to prescribe a particular element set if most of the elements are never used, or if their use cannot be relied upon across the entire collection." Following this assumption, in [Ochoa, Duval] the definition for the Completeness is presented as the degree to which the metadata record contains all the information needed to have an ideal representation of the described object. It is

clear that there are different ways for considering complete a metadata record by a users or a community.
Unfortunately this approach does not seem 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. A certification service has to avoid tailoring the assessment rules to bad practices.
In fact, there are several reasons that can determinate an empty value in a field. In [Guy and Powell, 2004], analyzing the quality of metadata in an Eprint archive, the authors have identified in the publication workflow the main issue. In fact, these repositories software are quite general purpose and require a certain degree of customization when are adopted by a Designed Community[37]. This customization concerns also the definition of which fields are mandatory, which values are expected and so forth. For instance, if the repository user interface allows you to skip the insertion of a value while it is considered mandatory or recommended by guidelines and standards adopted in a Designed Community, the probability of skipping it during the submission phases rises. The result is a very low Completeness score.
In summary, the Completeness dimension is function of the weight assigned to the field by the Designed Community according to standards and guidelines. The corrective actions to face the completeness occurrences are ranked according to the weight and the usage statistics.

Accuracy

In the Bruce and Hillman framework [Bruce, Hilmann, 2004] 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 value in the field are in the format expected

The same point of view is adopted by Stvilia [Stvilia, 2006l] when defines Accuracy/Validity dimension of the Intrinsic IQ as:"the extent to which the content information is legitimated or valid

[37] 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

according to some stable reference source such as a dictionary, standard schema and/or set of domain constraints and norms".
Following this point of view, the cases presented as a Consistency issue in [Ochoa, Duval] can be addressed by the Accuracy dimensions.
In [Ochoa, Duval], the authors identified three ways in which the logical consistency can be broken:
1) Instances include fields not defined in the standards,
2) Categorical fields that should only contain values from a fixed list, are filled with a non sanctioned value
3) The combination of values in categorical fields is not recommended by the standard definition.
Apart the first case, the last two cases refer to the value in a field that it is not expected by the standard definition. Thus, these classes of cases, according to our definition, fall into our Accuracy category.
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, is specified that the DC:subject has to assume the MIUR disciplinary sector values, while the DC:type field value has to be compliance with the MIME[IETF RFC 2045, 1996] definition, where an URI 38 is expected (DC:identifier), it is required a syntax correctness check.

In summary, we are in 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 on respect to the value expected by the standard.

Consistency

Some synonyms of Consistency referred to the data can be: compliance, non-contradictory, data reliability. In database domain, for example, if you need to change or delete a value that is linked to the others, the other fields must be updated or deleted, otherwise the data will result inconsistent. In fact, a task of the DBMS is to

[38] Uniform Resource Identifiers IEFT RFC 3986

assure a referential integrity[39] of the data. If there is no such control one would not know which of the different values is correct.
In [Stvilia, 2006l], inconsistency is considered internal if is referred to a single record or external if it is emerged among records. Moreover, Stvilia identifies two consistency problems: semantic and structural problems. From his point of view, they are measured by looking at data value on the conceptual level and data format on the structural level. The semantic consistency entails the degree of which the same data values or elements are used for delivering similar concepts in the description of a resource. A structural inconsistency concerns the extent to which the same structure of format is used for presenting similar data attributes and elements of a resource. One example is the different formats for encoding the date element such as dd-mm-yyyy or yyyy-mm-dd. Different formatting and use of different precision and scales for information elements result in structural inconsistency.
According to the Accuracy definition provided above, we consider these cases as an Accuracy issue since the data formats expected for each field are exactly what defined by the guidelines and standards. From the research perspective, the Consistency dimension has to address the logical error. In the 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 such as those that provide information to access to resources. In fact there is an accessibility issue when metadata retrieved does not allow the physical access to the digital content. If

[39] http://databases.about.com/cs/administration/g/refintegrity.htm

the metadata schema provides information for obtaining the resource, for example via URL, 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 reason 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.

In summary, the consistency issues emerge when the value in the field is formally compliance to the standard but is logically wrong. To this end, Consistency evaluation can be performed only if the Accuracy evaluation is passed positively. The Accuracy can be assessed in the Completeness evaluation is successfully passed.

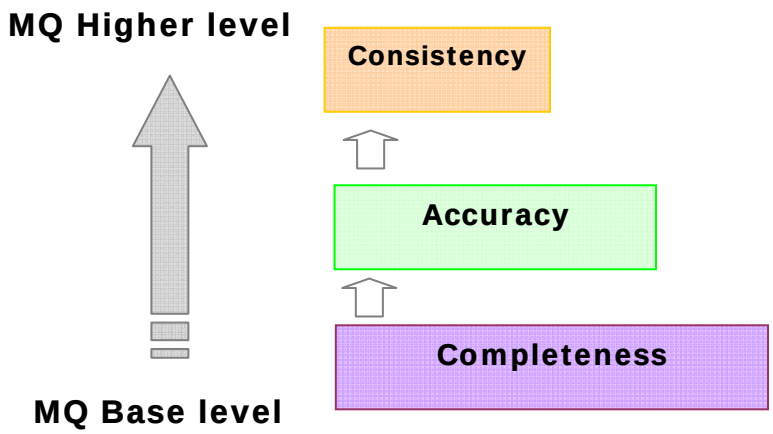


Figure 2 – Multi level MQ assessment

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 the then the number of field assessed during the Completeness evaluation where all field of the metadata set are processed. The same consideration is for Consistency box respect to the Accuracy one.
This is due to the fact that for some field it is really difficult to evaluate accuracy or consistency dimension with an automatic process. An example is the DC:coverage because there are not

defined specific rules and guidelines for its encoding, and if the value is arbitrary, its evaluation results impossible.

3.3 Low Level Metrics (LLM) definition

According to the GQM approach an early preparatory phase is needed to indentify the business goals. Form the point of view of this work, that phase is represented by the quality dimension definition. In fact, the business goal of the Metadata Quality Certification service is to assess the metadata records according to its quality model and provide a list of corrective actions to the institutions analyzed.

The Metadata Quality dimensions provided can be assessed at three levels: metadata field, metadata record, and community level. In particular the metadata field level foresees metrics that are able to evaluate the completeness, accuracy and so forth 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 for a defined community.

Completeness (COM)

Goal	
Purpose	To Evaluate
Issue	Completeness of
Object	IR metadata
Viewpoint	From the Standard and guidelines definition

Question – Q1 – Primitive
Which is the completeness of the i-th field in the IR for the community c and schema s?

Metrics	
ComplField$_i$	ComplField$_i$ **=** $\frac{\sum_{m=1}^{n\,\mathrm{Re}cord} f_m(field_i)}{n\,\mathrm{Re}cord}$ where: *field*$_i$ is the i-th field in the schema *m* is the m-th record *nRecord*: the total amount of the IR records and $f(x)=\begin{cases}0, \text{if the i - th field is empty}\\1, \text{otherwise}\end{cases}$ and nRecod is the number of records in the IR
Question – Q2	
which is the completeness score of records for the IR?	
Metrics	
ComplRecABS (Record completeness)	ComplRecABS = $\frac{\sum_{m=1}^{n\,\mathrm{Re}cord}\sum_{i=1}^{nField} f(field_i)}{n\,\mathrm{Re}cord}$ Where: *nField* is the number of the fields in the schema $\sum_{i=1}^{n} f(field_i)$ = the completeness of the single record Value range: from 0 to *nField* Where *nField* is the number of fields under investigation
Question – Q3	
Which is the completeness score weighed?	
Metrics	
ComplRecW	

	ComplRecW = $\frac{\sum_{m=1}^{n\,\mathrm{Re}cord} (\sum_{i=1}^{nField} f(field_i) \cdot w_i)_m}{n\,\mathrm{Re}cord}$ Where w_i is the i-th weight associated to the i-th field Value range: from 0 to $\sum_{i=1}^{nField} w_i$ Where *nField* is the number of fields under investigation

Accuracy (ACU)

Goal	
Purpose	**Evaluate**
Issue	**Accuracy of**
Object	**IR metadata**
Viewpoit	**From the Standard and guidelines definition**

Question – Q1	
Which is the Accuracy score of the i-th field	
Metrics	
AccurField$_i$	**AccurFields$_i$** = $\frac{\sum_{m=1}^{n\,\mathrm{Re}cord} f_i(field_i)_m}{n\,\mathrm{Re}cord}$ Where: $f_i(x)$ = is the accuracy function associated to the *i*-th field. $f(x) = \begin{cases} 0, \text{no problem founded} \\ 1, \text{if an accuracy issue was detected} \end{cases}$
Question – Q2	

Which is the Accuracy weighted score of the IR	
Metrics	
AccurRecW	**AccurRecW ==** $\frac{\sum_{m=1}^{n\mathrm{Record}} (\sum_{i=1}^{nField} f_i(field_i) \cdot w_i)_m}{n\mathrm{Record}}$ Where w_i is the weight associated to the i-th *field* of the schema Value range: from 0 to $\sum_{i=1}^{nField} w_i$ Where *nField* is the number of fields under investigation

Consistency (CON)

Goal	
Purpose	**Evaluate**
Issue	**Consistency of**
Object	**IR metadata**
Viewpoit	**From the Standard and guidelines definition**

Question – Q1	
Which is the Consistency score of the i-th field	
Metrics	
ConsField$_i$	**ConsFields$_i$** = $\frac{\sum_{m=1}^{n\mathrm{Record}} f_i(field_i)_m}{n\mathrm{Record}}$ Where: $f_i(x)$ = is the consistency function associated to the *i*-th field. $f(x) = \begin{cases} 0, \text{no problem founded} \\ 1, \text{if a consistency issue was detected} \end{cases}$
Question – Q2	
Which is the Consistency weighted score of the IR	
Metrics	

ConsRecW	**ConsRecW ==** $\frac{\sum_{m=1}^{n\,\mathrm{Re}cord} (\sum_{i=1}^{nField} f_i(field_i) \cdot w_i)_m}{n\,\mathrm{Re}cord}$ Where w_i is the weight associated to the i-th *field* of the schema Value range: from 0 to $\sum_{i=1}^{nField} w_i$ Where *nField* is the number of fields under investigation

Quality Score

Goal	
Purpose	**Estimate**
Issue	**The overall quality of**
Object	**The IR metadata records**
Viewpoit	**From the Standard and guidelines definition**

Question – Q1	
Which is the quality of the IR?	
Metrics	
MQC	To evaluate the quality index we use the weighted average. **WCompl=** $\frac{1}{\sigma^2_{Compl}}$ **; WAccur=** $\frac{1}{\sigma^2_{Accur}}$ **; WCons=** $\frac{1}{\sigma^2_{Cons}}$ **MQC=** $\frac{(\mathrm{ComplRecW} \cdot WCompl) + (Accur\,\mathrm{Re}cW \cdot WAccur) + (Cons\,\mathrm{Re}cW \cdot Wc}{WCompl + WAccur + WCons}$ **MQC =** $\frac{\sum_{i=1}^{n} x_i / \sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}$ Where n is the number of the dimensions addressed.

Chapter 4

Measurements plan

4.1 Introduction

David Zubrow in [Zubrow, 2007] proposes four objectives that a measurement process has to attain:
a) Characterize
To understand the current process, product, and environment
To provide baseline for future assessment

b) Evaluate
To determinate status so that project and process can be controlled
To assess the achievement

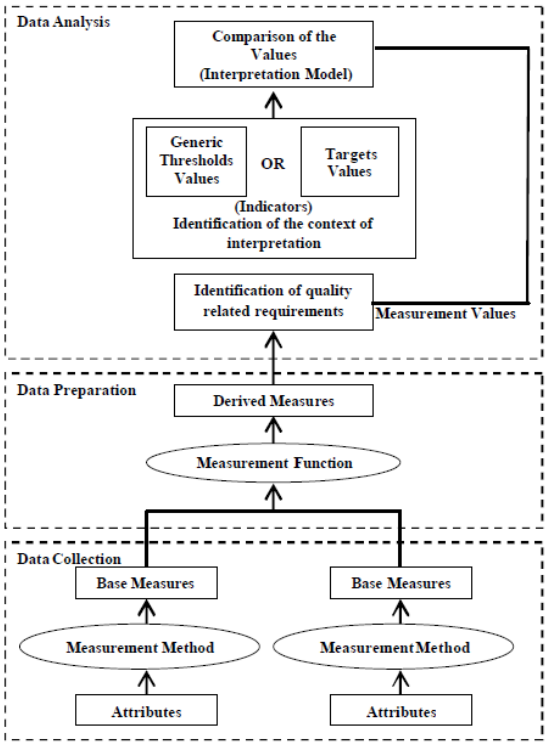
c) Predict
To understand the relationship between and among processes and products
To establish achievable goals for quality, costs and schedules

d) Improve
To identify root causes and opportunities for improvement
To track performance changes and compare baselines
To communicate reasons of improving

These objectives reflect the overall approach and aims of this work. For instance, the service provides as outcome of the metadata quality evaluation, a report to improve the metadata quality following a prioritised improvement actions.
In order to avoid the risk of getting overwhelmed with data, as outlined in [ISO/IEC 15939, 2002] and [Berander, Jönsson, 2006] one factor of defining successful measurement frameworks is to start small with the most important measurements and grow slowly as the organization matures, especially if measurements are being tried for the first time [ISO/IEC 15939, 2002]. At the beginning, in fact, is difficult understand which measurements are important, and

there is a risk that no measurements are collected and analyzed because it is not clear which ones to focus on (or that the "wrong" measurements are collected which is a waste of effort) [Fenton. Pfleeger, 1997][Berander, Jönsson, 2006]. Thus it is better to use a few number of useful measurements than none at all.

Within ISO 15939 (2002), is defined an information model that describe a measurement workflow highlighted the steps needed[ISO/IEC 15939, 2002]. The Figure shows that a specific measurement method is used to collect a base measure for a specific attribute. Then, the values of two or more base measures can be used within a computational formula (by means of a measurement function) to produce and construct a specific derived measure.
These derived measures are then used in the context of an analysis model to arrive at an indicator which is a value, and to interpret the indicator's value to explain the relationship between it and the information needed, in the language of the measurement user, to produce an Information Product for his Information Needs [ISO/IEC 15939, 2002].



ISO15939 Measurement workflow

4.2 Measurement plan

The table below reports the 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 extraction. More details on these tools are reported in the Prototype section. Because of the missing of a formal definition (free text) of some fields such as Creator or Contributor for assessing the Accuracy dimension, to these fields have not been associated any measurement functions, hence they will be not computed. Regarding the Consistency dimension the main issue is the complexity of the evaluation. In some case, the measurement provided in the table below like MIME comparison and broken link check remained as a proposal.
This plan aims to evaluate the metric not taking into account the number of the rules violated in the IR as defined in [Ochoa, Duval] because if the IR contains an high number of instance that are correct and only a few instance have made all the possible errors, the quality evaluation will be very low. Instead, we think that the use of the number of instance with problems reflects better the real status of the IR.

		Attributes		
DC field	**Rip**	**Completeness**	**Accuracy**	**Consistency**
dc:creator	S	Javscript Rule (at lease one instance)	NA	NA
dc.title	S	Javscript Rule (at lease one instance) Result: 0/1	Pear Language detect + Aspell Spelling check Result: 0/1 + list of wrong word	Semantic distance between the title and the article (TODO)
dc.subject	S	Javscript Rule (at lease one instance) Result: 0/1	Javascript Rule Comparison with the MIUR subjects list	NA

			Result: 0/1	
dc.date	N	Javacript Rule Result: 0/1	Isdate() - Yyyy - Yyyy-mm-dd - dd-mm.yyyy Result: 0/1	NA
dc.coverage	N	Javscript Rule Result: 0/1	NA	NA
dc:description	N	Javscript Rule Result: 0/1	Pear Language detect + Aspell Spelling check Result: 0/1 + list of wrong word	Semantic distance between the description and the article (TODO)
dc:relation	N	Javscript Rule Result: 0/1	NA	NA
dc.publisher	N	Javscript Rule Result: 0/1	NA	NA
dc:contributor	S	Javscript Rule (at lease one instance) Result: 0/1	NA	NA
dc:identifier	S	Javscript Rule (at lease one instance) Result: 0/1	Javascript rule for HTTP validator Result: 0/1	Javascript rule HTTP broken link check Result: 0/1
dc.language	N	Javscript Rule Result: 0/1	Javascript Rule for ISO 639-2/ ISO 639-1 Check Result: 0/1	Comparison between language value and language detected from the ojecter (TODO)
dc:type	N	Javscript Rule Result: 0/1	Javascript Rule Comparison with CRUI-DRIVER-MIUR object type definition	NA

			Result: 0/1	
dc.rights	S	Javscript Rule (at lease one instance) Result: 0/1	NA	NA
dc:format	S	Javscript Rule (at lease 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

Since some fields are repeatable (Rip= S), the Accuracy dimension for the field F is Accuracy (F) = $\frac{\sum f(i)}{n}$ where *f(i)* is the Accuracy evaluation function for the field F, *i* is the *i*-th instance of the field F and n is the total number of instance of the field F in a single metadata record. In the example below, the field dc:creator appears twice thus the *f(x)* associate to the field creator is applied two times. The result of the accuracy evaluation of the fields is divided for the total number of the instances (in this case 2). In this way the impact of each instance of a field in the field accuracy evaluation is $\frac{1}{n}$.

```
<dc:language>eng</dc:language>
<dc:creator>Berchum, Marnix</dc:creator>
<dc:creator>Rodrigues, Eloy</dc:creator>
<dc:contributor>Brown, John</dc:contributor>
```

The table below presents the mapping between the CRUI categories, MiUR categories and DRIVER v2.0 presented in the CRUI metadata guidelines document. These values are expected in the DC:type field for the measurement process.

CRUI	DRIVER v2.0	MiUR
Articolo in periodico	Article	Articolo su rivista
Contributo in libro	bookPart	Articolo su libro
Curatela	Book	Curatele
Libro	Book	Monografia
Brevetto	Patent	Brevetto
Tesi di dottorato	DoctoralThesis	Altro
Tesi magistrale	BachelorThesis	Altro

Tesi di master	MasterThesis	Altro
Intervento a convegno	ConferenceObject	Proceedings
Atto di convegno	ConferenceObject	Proceedings
Altro	Other	Altro
Recensione	Review	Articolo
Working paper	WorkingPaper	Altro

4.3 Assessing measurement validity

In the ISO/IEC 2520 Software and System Engineering – Software quality requirements and evaluation (SQuaRE – Quality measurement – Measurement reference model and guide[40] are illustrated which are the methods to demonstrate the validity of measures. In general these methods involve both a logical argument and statistical evidence. For instance the Lines of Code could as a measure of size has face validity because it is logically related to common notions of size. In many instances, simply documenting the rationale for the validity of a measure may be sufficient to ensure that the measure will yield meaningful results.
According to this assumption completeness is measured simply checking the presence/ filling of the metadata fields collected by OAI-PMH. The measurement result is an option yes/no in case the metadata field is empty/present or not in the set respect to the DC schema taken as reference for this research.
Moreover, this evaluation is weighted according to standards and guidelines. For the Accuracy and Completeness assessment the approach is the same.
Sometimes, statistical evidence of validity is required and can take several forms. Some examples of systematic variation are described below. According to the ISO/IEC 15939, the Repeatability and Reproducibility of the measurement done during this work can support this proof. In particular the Repeatability is referred to the degree to which the repeated use of the base measure in the same Organisational Unit following the same measurement method under the same conditions (e.g., tools, individuals performing the measurement) produces results that can be accepted as being identical . The Reproducibility refers to the degree to which the

[40] ISO/IEC JTC1/SC7/WG6 MEASUREMENT REFERENCE MODEL AND GUIDE http://cs.joensuu.fi/pages/pages/intra/saja/tSoft/FiSMA/fisma/paketti2003_1/25020MeasRefMode.pdf

repeated use of the measure in the same Organisational Unit following the same measurement method under different conditions (e.g., tools, individuals performing the measurement) produces results that can be accepted as being identical.
This work is based on an automatic computer based assessment on a specific dataset collected at T1 time. Thus, the repeatability and reproducibility of the measurements is an intrinsic characteristic of the overall research because is based on a fixed dataset where some defined criteria can be applied n times with the same results.

Third section

MQC service design and prototyping

Chapter 5

Metadata Quality Certification (MQC) service

5.1 Introduction

"Certification demonstrates to your customers, competitors, suppliers, staff and investors that you use industry-respected best practices."[41]

Certification provides benefits to Organizations, process improvement, employees, customers/user, and so forth. In fact, the Certification helps the organizations to demonstrate to stakeholders that their mission is running effectively, allows a better management controls, and increases the credibility of the institution[42]. Moreover, the process of achieving and maintaining the certification also helps ensure that the institution is continually improving and refining its activities, obtains greater employee awareness about quality and fewer problems with failures in service or product quality[43]. Certification can also improve overall performance, increasing productivity, remove uncertainty and makes it easier to satisfy user requirements.

In order to provide these benefits to the Open Access domain, this work aims to set up a Quality Certification Service for IR metadata. The objectives are three-fold:

a) to rise up the credibility and visibility of the open access resources empowering the user retrieval and access possibility b) to reduce the institutional cost for maintaining repositories with high quality metadata c) to support the standardization of the of the Open Archive pushing the institution to align their current practices to guidelines and recommendations.

The implementation of a quality certification service presents some critical issues such as the objectivity of the metrics and criteria, the authority and independence of the organization that manages the

[41] http://www.bsigroup.hk/en/Assessment-and-certification-services/Management-systems/At-a-glance/Benefits-of-certification/

[42] http://www.qualitygurus.com/download/ISO9001BenefitsOfISO9001Certification.pdf

[43] http://www.qualitygurus.com/download/ISO9001BenefitsOfISO9001Certification.pdf

service, and so forth. As defined in CMMI [Forneser, Brurteau, Shrum], the Process and Product Quality Assurance (PPQA) is an objective insight into process and associated work products provided to staff and management. The PPQA defines the following activities:

- Objectively evaluating performed process and work products against applicable process description, standards and procedures;
- Identifying and documenting noncompliance issues
- Providing feedback to project staff and managers on the results of quality assurance activities
- Ensuring that noncompliance issues are addressed

The Objectivity in process and products quality assurance evaluations is critical to the success of the process. Objectivity is achieved by both independence and the use of criteria [Forneser, Brurteau, Shrum]. Examples of way to perform objective evaluation include the following:

- Formal audit by organizationally separate quality assurance organizations
- Peer review, which may be performed at various level of formality
- The ISO9000 certification provided by a independent third-parts that certify your workflow, products, process, etc. following a standardized assessment.

It is clear that if the quality assurance is embedded in the process, it is hard to obtain a reliable quality assessment because the results can be manipulated. Thus, as defined previously, this work tackles these issues:

a) Defining a Metadata Quality Framework starting from the common standards such as ISO 9124 and ISO 25000, and taking into account the metadata requirements defined by NISO and FRBR model. Moreover, the major metadata quality frameworks have been analyzed and revised.
b) Defining a number of related metrics following the well known GQM approach.

c) Defining the measurement process according to ISO measurements process definition and declaring the measurement criteria.
d) Defining a QP according to IFLA-FRBR, CRUI guidelines and User Community.
e) Indentifying a trust 3 rd party for running the service.

In particular a metadata certification service could be supported the legal deposit service provided by the consortium of national legal deposit.[44]
The aim is to assure objectivity and independence of the results through an independent reporting channel and at the same time obtain a compliance evaluation against shared QP.

5.2 Repository Certification initiatives

The constant growth of the digital objects present in the Web without a clear definition of their provenance, authenticity, authority, etc, affects the credibility of Internet as a reliable channel for disseminating a retrieving cultural heritage and scientific contents. In fact, the final user requires evidences before reusing these kind of resources because a wrong information could hit research results at all levels. This situation pushes the repositories to provide the evidence that their content have all characteristics required for being reused safely. Thus several certification initiatives are running for testing the credibility of the repositories against, for instance, preservation capability, risk resilience, or trustworthiness in general. This paragraph reports the main initiatives in the field that are used as inspiration for this work.

Data Asset Framework The Data Asset Framework focuses on uncovering researchers' data needs and concerns. It was created in the UK for its Higher Education Institutions to help them assess their data holdings and ensure appropriate data management practices are in place.

[44] The consortium is composed by National Library of Florence, Rome, Venice and Fondazione Rinascimento Digitale www.depositolegale.it

Data Seal of Approval [45](DSA) The Data Seal of Approval is intended to certify Data Archives who house research data within scientific and scholarly research domains. Archives must meet sixteen guidelines to be certified. The certification is granted by an Approval Board. The board includes members who are employed in a variety of international data archives. The archive, once certified, will be permitted to display the DSA logo on its homepage, and in other locations relevant to its communication.

Digital Asset Assessment Tool [46](DAAT): Project Digital Asset Assessment Tool (DAAT) Project is the University of London Computer Center's guide to the risk factors that may affect the survival of digital assets.

DIN 31644: Germany's DIN Standards Committee on Information and Documentation (NABD) is responsible for the standardization of practices relating to libraries, documentation and information centers, indexing and abstracting services, archives, museums, information science and publishing industries. The DIN 31644standard is a set of criteria that define standardized requirements for the setup and management of digital archives. The DINI [47]certification criteria was an initiative of the Deutsche Initiative für Netzwerkinformation (German Initiative for Networked Information).

DRAMBORA [48]**risk assessment:** The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) is a toolkit for use by repository administrators to assess the risks to their digital archiving systems.

ISO 2146 Project: ISO 2146 [49](Registry Services for Libraries and Related Organisations) is an international standard currently under development by ISO TC46 SC4 WG7 to operate as a framework for building registry services for libraries and related organisations. It takes the form of an information model that identifies the objects and data elements needed for the collaborative construction of

[45] http://datasealofapproval.org
[46] http://www.data-audit.eu/tool.html
[47] http://www.dini.de/
[48] http://www.repositoryaudit.eu/
[49] http://www.nla.gov.au/wgroups/ISO2146/

registries of all types. It is not bound to any specific protocol or data schema. The aim is to be as abstract as possible, in order to facilitate a shared understanding of the common processes involved, across multiple communities of practice. http://www.nla.gov.au/wgroups/ISO2146/

NESTOR Catalogue: NESTOR is the German agency assigned the task of providing libraries, archives and museums information and training on digital preservation. Among other digital preservation activities, NESTOR Working Groups develop standards for digital preservation. These standards are adapted by DIN (see above) as national standards: Catalog of Criteria for Trusted Digital Repositories [50]was published in 2007 by a NESTOR working group, Version II was introduced in 2010. They have also produced the Catalogue of Criteria for Assessing the Trustworthiness of PI (Persistent identifiers) Systems

P**lanning Tool for Trusted Electronic Repositories** [51](PLATTER) developed a tool, called the Repository Planning Checklist and Guidance in 2006 that is useful for digital repository planning.

SHAMAN Assessment Framework [52]The EU-funded SHAMAN (Sustaining Heritage Access through Multivalent Archiving) project is developing an integrated preservation framework using grid-technologies for distributed networks of digital preservation systems, for managing the storage, access, presentation, and manipulation of digital objects over time.

Trustworthy Information Systems Handbook [53]This handbook, developed in 2002, provides a set of criteria to establish the trustworthiness of government information systems.

5.3 Quality Certification Service scenarios

In order to face the open issues described above, we have designed the system following the Scenario Based Design (SBD) principle

[50] http://www.langzeitarchivierung.de/eng/schwerpunkte/standardisierung.htm
[51] http://www.digitalpreservationeurope.eu/platter.pdf
[52] http://shaman-ip.eu/shaman/document
[53] http://www.mnhs.org/preserve/records/tis/tableofcontents.htm

[Carroll, 1995]. Scenarios are a vocabulary for coordinating the central tasks of system development, understanding people's needs, envisioning new activities and technologies, designing effective systems and software, and drawing general lessons from systems as they are developed and used. The basic argument behind scenario-based methods is that descriptions of people using technology are essential in discussing and analyzing how the technology is (or could be) used to reshape their activities. A secondary advantage is that scenario descriptions can be created before a system is built and its impacts felt.
According to [Carroll, 1995], there are three different use modalities of the scenario:

1) to analyse of activity for structuring data collected from observation of the user tasks,
2) to prototype for envisioning the future task and stimulate the design process, and
3) to evaluate for testing the existing solutions.

In this scenario we have worked on all three levels. We have analysed the user activities (to manage IR) and the present issues. Then it is provided an envisioning scenario with a focus on the specific users (university), what are their goals (disseminate AO resources), what activities have to do to achieve the objectives (correct errors on metadata records) , and the context (University Library Dept.) for driving the technology integration [Donatelli, et, al, 2005].

Analyse and Evaluation scenario
The University A is one of the biggest universities in the country with thousands of students and a several faculties. It has set up an Institutional Open Archive managed by library department open to all university scientific results and adopted an institutional policy that defines as mandatory the deposit of research results in the Open Archive since they are funded with public money.
During the set up of the service, it has not been established any particular quality controls or formal audits for the workflow results. Moreover, the lack of resources, awareness about the importance of such archive for the institution and the adoption of inappropriate software tools [Guy and Powell, 2004], has privileged a fast self

published process of the research products with basic and low accurate metadata.
In any case the publication rate is high because of the number of the researchers that work in the institution.

Wishing to refine the evaluation criteria for the universities, the Italian Ministry of Research and University (MiUR) decides to include the institutional repositories in the research evaluation process, in order to rank the university properly and assign them new funds. Moreover, according to the same aims, the MiUR decides to compare the impact results taking into account not only the Impact Factor (IF) but other bibliometric indexes based on web ranking, citation, etc.

At this point, the people employed at University A , from manager to researchers, understand the quality of metadata associate to resources is crucial. In fact, the research products retrieved and accessed easily by the users, are more cited then the others.
Unfortunately the accessing rate is very low respect to the amount of resource stored in the archive and the risk of a bad evaluation for the MiUR is concrete. Because of the lack of resources and a low number of personnel in the library department, the manager of university A decide to start a general metadata assessment of the repository records assigned this duty to some students.
The work has not produced any significant results. In fact, some missed information are more clear to librarian or archivist (e.g. the ISSN number), others can be filled or corrected only by the paper authors (e.g. the Author name), and so forth. Moreover, to detect these problems, each record has to be controlled. Since the repository of university A contains over 150.000 records, it is impossible to estimate how long this process will take. Finally, since there are not tools able to support and optimize these effort (for instance, using a functionality that select most critical records first) once the assessment is lunched, the increment of the quality level reached during the work is totally unpredictable. The problem is not knowing which is the right proportion between effort provided and quality reached. It is a crucial management problem that could have broader implications.

These scenarios force the institutions to spent moneys and time to revised all metadata records looking for errors without ongoing

activities controls and possibility to set deadline and targets. Another problem emerges from the scenario analysis. In fact, for IR that collects all the production of very small research institutions and universities, where the publication rate is very low, maintaining a quality control on the metadata ingested could be not a problem. Instead, for medium and big institution, this is impossible without automatic tools that have to make a check periodically, with close intervals. This is the only way to maintain under control the quality of metadata during the IR activities, avoiding a general quality fall.
Finally, detecting metadata errors can reveal other problems at different level of the institution workflow. For instance, if a metadata field tents to be empty in ever new ingestion along the lifetime of the IR, the problems could be related to collecting procedures established by institution that are not clear enough for the final users, or the user interface of the system is not well designed, and so on.

Envisioning scenario

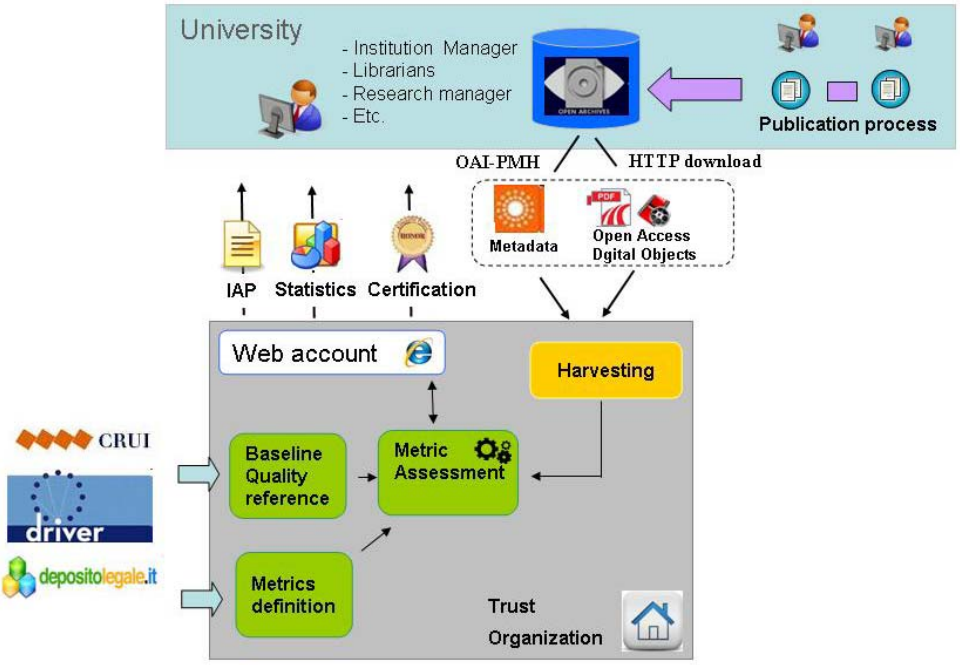
To tackle the situation, the University A manager decides to undertake a metadata certification process to promote to the stakeholders (MiUR, Private sectors, foundations, science communities, etc.) the quality of the institutions through a better dissemination of its research products.

After an "offline" agreement between the University A and Certification Authority, the library department registers their open archive to the MQC service. Through the university account, the library department can select the level of service required, monitoring the state of its repositories, manage the reports, and so forth. They can be free to decide also which metadata schema managed by the repository has to be assessed. The process starts to collect all metadata form the repository. If It is required a finer evaluation also digital objects can be collected.
Then the system evaluate all the metrics defined and as a result provides a report of corrective actions ranked on the base of their impact on the resource retrievability and accessibility.
The Metadata Quality Certificate is released when the global quality of the metadata falls into a defined range. The certificate validity is function of the publication rate and the index of the metadata quality deterioration after each publication.

In this way, the University manager can manage the efforts on the base of the quality target to be reached by a certain date. Moreover, the certificate validity will be longer in relation to the quality of the publication workflow.

5.4 MQC Service functionalities

MQC service has to provide an objective ad transparent metadata evaluation service, estimating metrics, reporting a series of corrective actions and releasing the MQ Certificate if the criteria comply with threshold and criteria. The figure here below represents the main elements o the service, which are the relations with the institutions (e.g. universities) ad reference community.



According to the metadata requirements and the scenarios presented above the MCQ service should allow:

a) the registration of Open Archive to the MQC service inserting the OAI-PMH URL of the IR,
b) the harvesting via OAI-PMH of metadata records of the IR. If required the harvesting can be extended to each of metadata schema managed (DC, MPEG21, METS, MAG, etc.),
c) the evaluation a number of qualitative and quantitative metrics that contribute to definition of quality indicators. In

particular it is possible calculate metrics to different metadata set separately. For instance it is possible to have DC with quality x, MPEG21 with quality y, etc.),

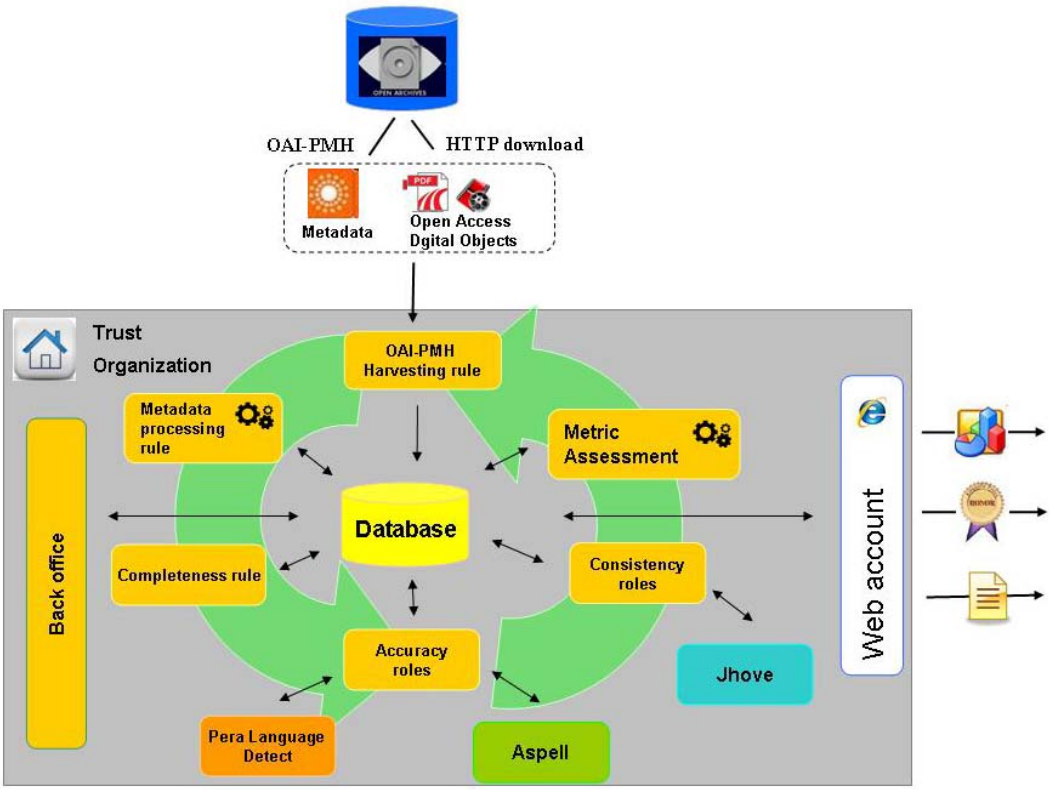
d) the possibilities to define more metrics based on community domain whereof the IR under evaluation belongs to (humanistic, physic, informatics, etc.),
e) the reporting of the evaluation result to the institution. This report could have different level of detail in relation to the type of service required by institution. The evaluation report is a guide for correcting the errors and lacks, with the objective to rise up the quality level of the IR,
f) the releasing of the MQ Certificate, if the quality level of the IR is above a specific threshold. This metadata quality certification has a defined temporal validity calculated on the base of the submission rate and the quality deterioration index. In fact the certification is related to a "snapshot" of the repository status at that moment,
g) the service can manage the historical data of each evaluation in order to analyze quality trends of that IR,
h) the service assure the security and confidentiality of the metadata harvested. If required, a backup of metadata can be maintained to understand quality evolutions at record level, and
i) the service is completely automated and the waiting time to obtain the evaluation report is only dependent on the harvesting and elaboration processes.

Chapter 6

Prototype Implementation

6.1 System architecture

The MQC service prototype implements a number of GRID rules that identify the steps of the assessment. The process start 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. The second step is performed by the metadata processing rule. This rule extract each single field form the metadata table and populate a table with rdf-like tripe and each row represent a field. Then the rules for completeness assessment can be lunched. After that, the accuracy can be assessed for each field through a proper evaluation rule. These rules require the 3-rd party applications. The next step is addressed by the consistency rule. This rule can be lunched only on the field that have passed positively the completeness and the accuracy evaluation. Finally, the metric assessment, calculates Average, variance and the MQ index. These results are presented through a web application where the user can interact with the system (setting the type of certification, level of details, metadata set to be evaluate, etc.).



Axmedis overview

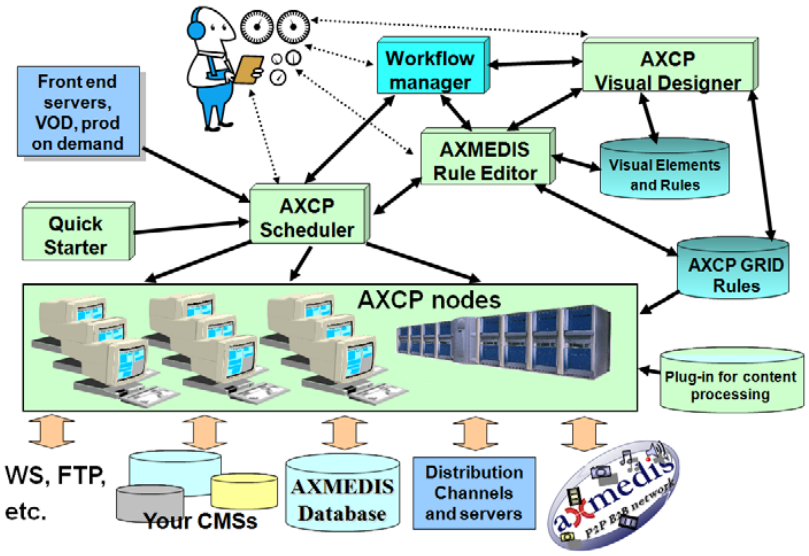
This MQC service is based on AXMEDIS framework[54], an open source infrastructure that allows massive harvesting, metadata processing and evaluation, automatic periodic quality monitoring, etc.

AXCP grid backoffice

The MQUA service is based on AXCP tool that can manage parallel executions of processes (called rules) allocated on one or more computers/nodes. The rules are managed by a central scheduler and are formalized in extended JavaScript (Bellini, Bruno, Nesi, 2006). 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

[54] AXMEDIS EU-project: Automating content of Cross Media Content for Multichannel Distribution http://www.axmedis.org

by means of WEB services (see Figure 3). 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.

Grid approach for harvesting

The solution approach is based on OAI-PMH protocol (see Annex 'A'), 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. It is well-known that web services are also a computing technique for systematically disseminating XML contents, but when the global amount of data

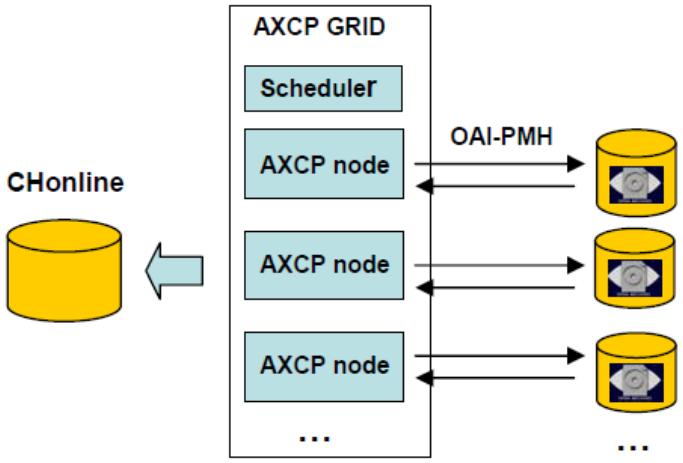
increases, some problems come out. According to OAI-PMH protocol, Guidelines for Harvesting Implements [Lagoze, Van de Sompel, Nelson, 2002] and OA implementation tutorial, a client may put a request to OAI server to ask for the stored content descriptors. Answers are related to the accessible records, and adopted formats. The OAIPMH protocol provides a list of discrete entities (metadata records) by XML stream. In many cases, these lists may be large and it may be practical to partition them among a series of requests and responses. In fact, the repository replies to a list request with an incomplete list and resumption Token. In order to get responses as much as possible from the list of the OAs considered, the harvester has been performed more requests with resumption Token as arguments. The complete list then consists of the concatenation of the incomplete lists from the sequence of requests, known as a list request sequence [Lagoze, Van de Sompel, 2002].
Moreover, in the current version of the OAI-PMH protocol a 'verb' to obtain the number of the records that we are going to harvest is not defined. Thus it is impossible to estimate a priori the duration of the process in terms of counted metadata sets. It is clear that the number of records included in a incomplete list (or page) affects the harvesting performance. In some cases this number was only one, and yet the harvester had to perform requests as many as the records in the archive. The harvesting performance also depends on response delay that is related to the network bandwidth and machine performance used by the connected Open Archive. In some cases, this time was greater than 15s for each request. In order to cope with the complexity, a parallel solution has been set up and used as described in the next subsection.

GRID based metadata harvesting architecture

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)[55]. The

[55] AXMEDIS EU-project: Automating content of Cross Media Content for Multichannel Distribution http://www.axmedis.org

computational solution has been implemented by realizing a parallel processing algorithm written in AXCP Extended JavaScript [Bellini,, Bruno, Nesi, 2009]. 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 can be periodically (or on demand) scheduled with an interval, for instance, of 1 minute from each running on a single node and the successive one. 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 called CHonline. In figure below, a schema of the architecture is shown. Each GRID node executes an identical autonomous harvesting rule that collects metadata from an Open Archive and populates the database according to the general status also collected into the 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.



Grid architecture for massive OA harvesting
via OAI-PMH protocol on grid infrastructure

GRID-based harvesting workflow

This paragraph describes the grid base harvesting algorithm and workflow. Figure 2 shows a schema representing the consecutive

steps performed by the harvesting rules on the grid. Before performing the effective harvesting of the single records, two preparatory steps are needed: (i) to get the repositories information; (ii) to get the metadata sets available for each repository. These two steps are performed into the grid with specific aperiodic/on-demand rules. During the first step a rule for getting the repository list from http://www.openarchives.org/Register/ListFriends website is launched. This rule parses the XML list of OA repositories baseURLs and populates the repository table of database. For example, a segment of the repository list is as follows:

[…]

<baseURL id="UOV.es">http://www.tdr.cesca.es/TDR_UOV/NDLTDOAI/oai.pl</baseURL>

<baseURL>http://diglib.cib.unibo.it/oai/oai2.php</baseURL>

<baseURL>http://docinsa.insa-lyon.fr/oai/oai2.php</baseURL>

[…]

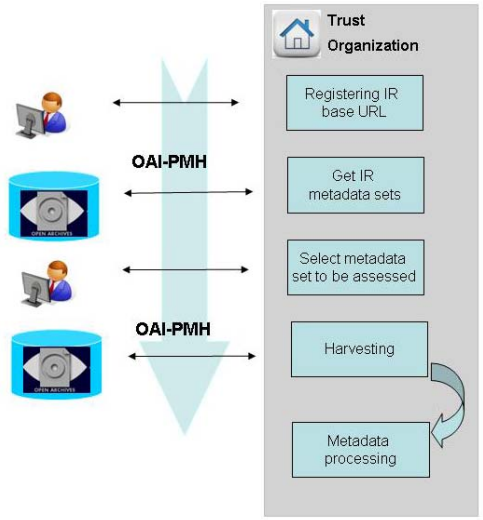


Figure .. – Algorithm for harvesting

The repositories are identified with a <baseURL> field filled in with the URL of repository OAI-PMH interface and the repository ID. This rule may be also periodically scheduled for checking the availability of new repositories added to the list that have to be harvested by the system. Once the repository list is obtained, the second step has to determine if the OA is active and which service may provide. To this end, a dedicated second rule is activated to both verify the

activity of the OA and retrieve the metadata formats available by using the ListMetadataFormats verb of OAI-PMH. A repository is set 'not available' if it does not provide any response, so that at the next round it can be tried again. In fact, it may happen that a repository may be offline for some reasons. Therefore, each single OA provides the list of metadata according to the following example.

http://baseURL/request?verb=ListMetadataFormat
[
<ListMetadataFormats>
<metadataFormat>
<metadataPrefix>oai_dc</metadataPrefix>
<schema> http://www.openarchives.org/OAI/2.0/oai_dc.xsd
</schema>
<metadataNamespace>http://www.openarchives.org/OAI/2.0/o
ai_dc/</metadataNamespace>
</metadataFormat>

The list of metadata sets of each repository is stored in the metadata formats table of database. It should be noted that if a metadata format is declared as being supported by an OA, this does not mean that it is available for all the items in the repository.
OA Harvesting. The harvesting rule gets access to the status table in database to obtain the first not processed archive/metadata-set and it starts with its crawling. Moreover, the harvesting rule parses the XML response, it extracts only the metadata information and it saves it in a single database field/chunk as a string. The harvesting rule is designed to harvest the records only from one repository managing the resumption Token. This approach is meant to reduce the rule time activity, but there are some cases where a rule could stay alive for hours (for instance if there are a lot of records to harvest and the OAI request has provided a short number of records).

6.3 OAI-PMH architecture

The Open Archive Initiative (OAI) [5] consists of a technical and organisational framework designed to facilitate the discovery of content stored in distributed archives such as e-print. It makes easy-to-implement technical recommendations for archives that –

when implemented – will allow data from e-print archives to become widely available via its inclusion in a variety of end-user services such as recommendation services, services for inter-linking documents, etc

The OAI architecture identifies two logical roles: "Data Providers" and "Service Providers". Data Providers deal with both the deposit and publication of resources in a repository and they "expose" for collecting the metadata about resources in the repository. They are the creators and keepers of the metadata and repositories of resources. At present many institutions have implemented the OAI Data Provider, thus choosing the following repository software: Dspace [56], Fedora [57], Eprints[58], Greenstone [59], etc.

Service Providers use the OAI-PMH interfaces of the Data Providers to collect and store their metadata. They use the collected metadata for the purpose of providing one or more services across all the data. The types of services, which may be offered, include a search interface, peer-review system, etc. The key architectural shift was to move away from only supporting human end-user interfaces for each repository, in favour of supporting both human end-user interfaces and machine interfaces for collecting.

OAI-PMH requests must be submitted using either the HTTP GET or POST methods. POST has the advantage of imposing no limitations on the length of arguments. There is a single base URL for all requests. The base URL specifies the Internet host and port, and optionally a path, of an HTTP server acting as a repository. Repositories expose their base URL as the value of the baseURL element in the Identify response. Note that the composition of any path is determined by the configuration of the repository's HTTP server.

In addition to the base URL, all requests consist of a list of keyword arguments, which take the form of verb=value pairs. Arguments

[56] DSpace <http://www.dspace.org>

[57] Fedora <http://www.fedora.info>

[58] EPrints for Digital Repositories <http://www.eprints.org>

[59] Greenstone University of Waikato <http://www.greenstone.org>

may appear in any order and multiple arguments must be separated by ampersands [&]. Each OAI-PMH request must have at least one verb=value pair that specifies the OAI-PMH request issued by the harvester.

```
<simpleType name="verbType">
  <restriction base="string">
    <enumeration value="Identify"/>
    <enumeration value="ListMetadataFormats"/>
    <enumeration value="ListSets"/>
    <enumeration value="GetRecord"/>
    <enumeration value="ListIdentifiers"/>
    <enumeration value="ListRecords"/>
  </restriction>
</simpleType>
```

Examples

Request:

List the records expressed in oai_rfc1807 metadata format, that have been added or modified since January 15, 1998 in the hep subset of the physics set [URL shown without encoding for better readability].

```
http://an.oa.org/OAI-script?
      verb=ListRecords&from=1998-01-
15&set=physics:hep&metadataPrefix=oai_rfc1807
```

Response:

Two records are returned:

* The first record is expressed in the oai_rfc1807 metadata. This record also has an about part, and the item from which it was disseminated belongs to two sets (physics:hep and math).
* The second has a header with a status="deleted" attribute (and therefore no metadata part).

Note: The reply only includes records for those items from which metadata in oai_rfc1807 can be disseminated. No records are returned for those items that fit the from, until, and set arguments but from which the specified format can not be disseminated.

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
        http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
 <responseDate>2002-06-01T19:20:30Z</responseDate>
 <request verb="ListRecords" from="1998-01-15"
         set="physics:hep"
         metadataPrefix="oai_rfc1807">
         http://an.oa.org/OAI-script</request>
 <ListRecords>
  <record>
   <header>
    <identifier>oai:arXiv.org:hep-th/9901001</identifier>
    <datestamp>1999-12-25</datestamp>
    <setSpec>physics:hep</setSpec>
    <setSpec>math</setSpec>
   </header>
   <metadata>
    <rfc1807 xmlns=
      "http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation=
      "http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1807.txt
       http://www.openarchives.org/OAI/1.1/rfc1807.xsd">
       <bib-version>v2</bib-version>
       <id>hep-th/9901001</id>
       <entry>January 1, 1999</entry>
       <title>Investigations of Radioactivity</title>
       <author>Ernest Rutherford</author>
       <date>March 30, 1999</date>
    </rfc1807>
   </metadata>
   <about>
    <oai_dc:dc

xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
```

```
        http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
       <dc:publisher>Los Alamos arXiv</dc:publisher>
       <dc:rights>Metadata may be used without restrictions as long
as
         the oai identifier remains attached to it.</dc:rights>
     </oai_dc:dc>
   </about>
  </record>
  <record>
   <header status="deleted">
    <identifier>oai:arXiv.org:hep-th/9901007</identifier>
    <datestamp>1999-12-21</datestamp>
   </header>
  </record>
 </ListRecords>
</OAI-PMH>
```

6.4 Metadata processing

The metadata harvesting is the first step to collect data and per se it not sufficient to evaluate the quality of metadata implementation. In fact, it is not possible to extract specific metadata values that are related to a specific argument. Moreover the high number of implemented different metadata sets requires a tool for processing them in order to get the single metadata element.

Moreover, 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[60] 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.

[60] W3C- RDF/XML Syntax Specification http://www.w3.org/TR/REC-rdf-syntax/,

6.5 3rd Party Tools

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

The concept of representation format, or type, permeates all technical areas of digital repositories. Policy and processing decisions regarding object ingest, storage, access, and preservation are frequently conditioned on a per-format basis. In order to achieve necessary operational efficiencies, repositories need to be able to automate these procedures to the fullest extent possible.

JSTOR and the Harvard University Library are collaborating on a project to develop an extensible framework for format validation:

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; in other words, it answers the question: "I have a digital object; what format is it?"
- Format *validation* is the process of determining the level of compliance of a digital object to the specification for its purported format, e.g.: "I have an object purportedly of format *F*; is it?"

  Format validation conformance is determined at two levels: *well-formedness* and *validity*.

  1. A digital object is well-formed if it meets the purely syntactic requirements for its format.
  2. An object is valid if it is well-formed and it meets additional semantic-level requirements.

  For example, a TIFF object is well-formed if it starts with an 8 byte header followed by a sequence of Image File Directories (IFDs), each composed of a 2 byte entry count and a series of 8 byte tagged entries. The object is valid if it meets certain

additional semantic-level rules, such as that an RGB file must have at least three sample values per pixel.

- Format *characterization* is the process of determining the format-specific significant properties of an object of a given format, e.g.: "I have an object of format *F*; what are its salient properties?"

  The set of characteristics reported by JHOVE about a digital object is known as the object's *representation information*, a concept introduced by the Open Archival Information System (OAIS) reference model [ISO/IEC 14721]. The standard representation information reported by JHOVE includes: file pathname or URI, last modification date, byte size, format, format version, MIME type, format profiles, and optionally, CRC32, MD5, and SHA-1 checksums [CRC32, MD5, SHA-1]. Additional media type-specific representation information is consistent with the NISO Z39.87 Data Dictionary for digital still images and the draft AES metadata standard for digital audio.

Identification, validation, and characterization actions are frequently necessary during routine operation of digital repositories and for digital preservation activities. These actions are performed by *modules*. The output from JHOVE is controlled by *output handlers*. JHOVE uses an extensible plug-in architecture; it can be configured at the time of its invocation to include whatever specific format modules and output handlers that are desired. The initial release of JHOVE includes modules for arbitrary byte streams, ASCII and UTF-8 encoded text, GIF, JPEG2000, and JPEG, and TIFF images, AIFF and WAVE audio, PDF, HTML, and XML; and text and XML output handlers.

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. Its main feature is that it does a superior job of suggesting possible replacements for a misspelled word than just about any other spell checker out there for the English language. Unlike Ispell, Aspell can also easily check documents in

UTF-8 without having to use a special dictionary. Aspell will also do its best to respect the current locale setting. Other advantages over Ispell include support for using multiple dictionaries at once and intelligently handling personal dictionaries when more than one Aspell process is open at once

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.

Chapter 7

OA Repository assessment

6.1 Assessment results

According to the above described solution for massive OA inspection and metadata harvesting, a set of metrics and considerations has been performed. They may be used to evaluate the implementation of OA as an effective tool for disseminating scientific works via OAI-PMH service protocol.
A champion of 9 IR was randomly selected. The unique selection requirements is their compliant with the OAI-PMH protocol.

UnipiEprint - University of Pisa

UnipiEprints is an institutional repository where you can deposit through the auto-archive process and preserve scientific contributions published by the teaching staff and researchers at the University of Pisa

BaseURL: http://eprints.adm.unipi.it/cgi/oai2
Number of records: 465
Last harvesting: 2011-12-13
Quality Score: 267259,7

Completeness	
Average	0,765195699
Standard Deviation	0,022313557
Variance	0,000497895
Minimum	0,658
Maximum	0,84
Livel of confidence(95,0%)	0,002033409

Accuracy	
Average	0,450954839
Standard Deviation	0,074540143
Variance	0,005556233

Minimum	0,282
Maximum	0,675
Level of confidence (95,0%)	0,006792758

The chart represent for each field the level of Completeness in the repository. The results shows in general that all fields required by the repository system seem filled.
In fact only few records have the field Contributor with a value and any 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.

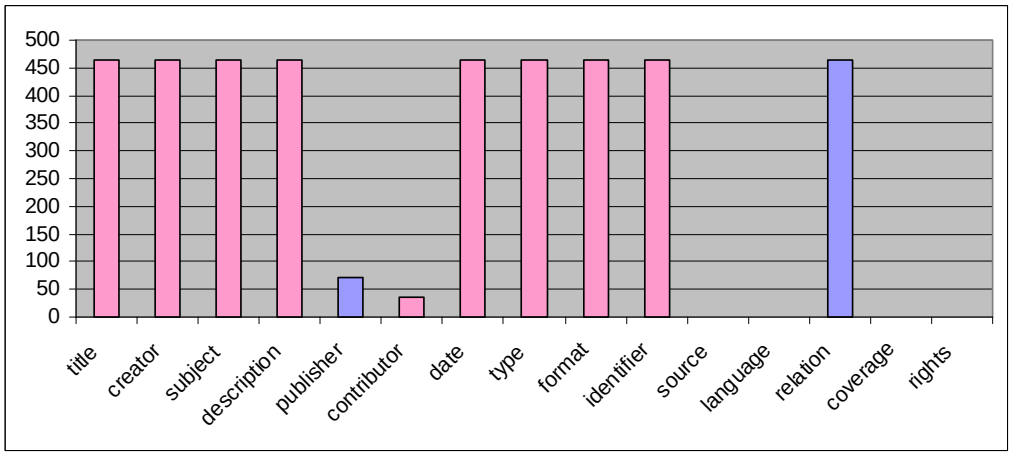


Chart 2 - Fields Completeness for IR of University of Pisa

The Accuracy chart shows the level of accuracy of each field weighted with the MQC weights. This chart show that the field description and title are those less accurate. This might be due to the type of the field. In fact the value expected is a free text and 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.
Instead, the field where authority files, fixed lists of values are defined tend to be fully accurate. In this case the subject field is filled with the value presented in the MIUR subject list. The Identifier is a repeatable field in a record where some instances are accurate and others are not.
For instance the following use of the field Identifier is out of the Accuracy rule, thus, in this research it is considered not accurate.

<DC:Identifier>Aria, Giorgio and Shou, Zhang and Botta, Roberto and Giuliotti, Lorella and Rota, Alessandra (2004) Trans-

vaginal echographic approach to early pregnancy diagnosis in small ruminants. Annali della FacoltÃ di Medicina veterinaria, LVII/2 . pp. 35-42. ISSN 0365-4729</DC:Identifier>

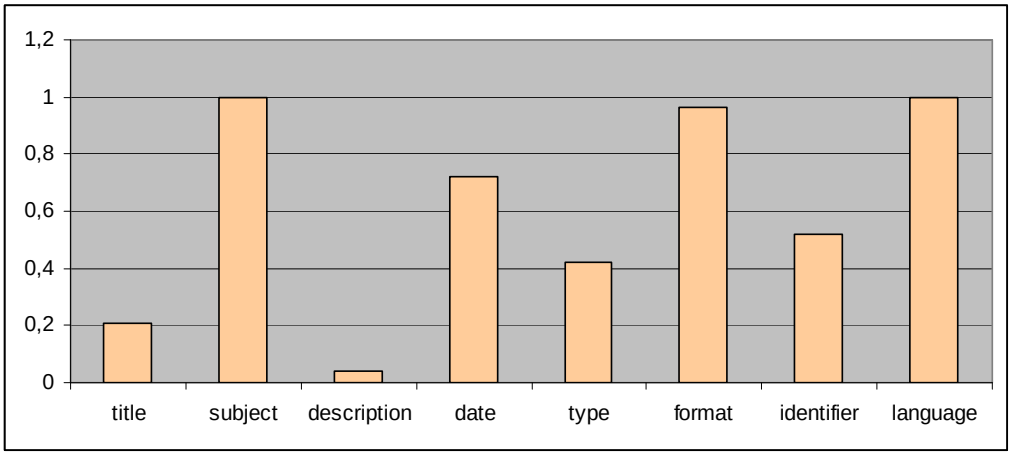


Chart 3 - Fields Accuracy for IR University of Pisa

The chart represents the monthly distribution of the Average of the Completeness.

This chart might also represent the reliability of the submission process. In other words, if the level of completeness distribution is the same during the time, this might be an index that a reliable workflow is in place because you have a standardized outcome independently of its quality result.

On the other hand, many oscillations in the completeness level might be due to internal or external factors respect to the institution. In fact the completeness can be the result of a mix of factors such as an usable interface (internal), clear policies (internal/external), trained staff (internal), number of submissions (external), and so forth.

Hence, the chart shows that there are some oscillations but they do not affect the general level of completeness because the overall Average is over the 75%. The Average of the Accuracy instead, is under the 45% with some oscillations without any correlation (-0,086908075) with the Completeness line.

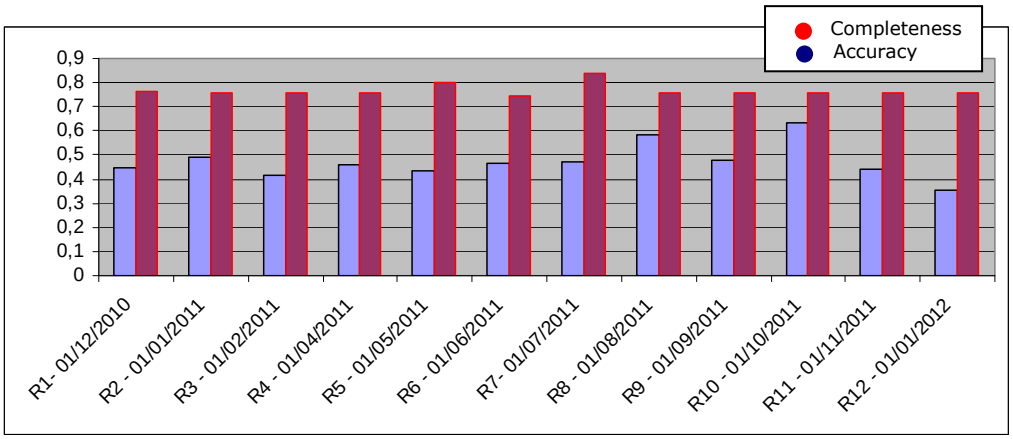


Chart 4 - Time chart of Accuracy and Completeness for IR of University of Pisa

This chart reports the distribution of the submission rate in the IR. The results shows a substantial underuse level. The peak represents the start-up of the IR, a massive submission to populate it.

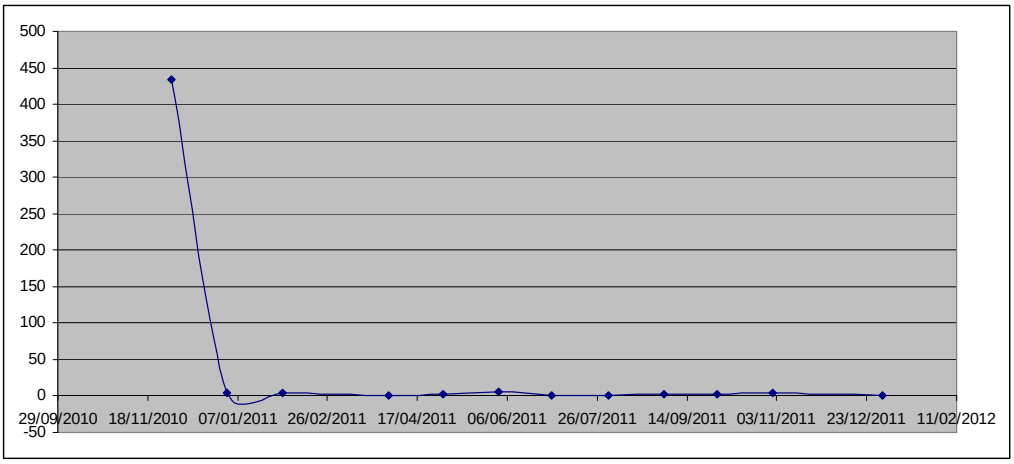


Chart 5 - IR Submission rate University of Pisa

In this chart is put in evidence the underuse of the IR. In Fact the submission does not exceed the number of 6 in a Month.

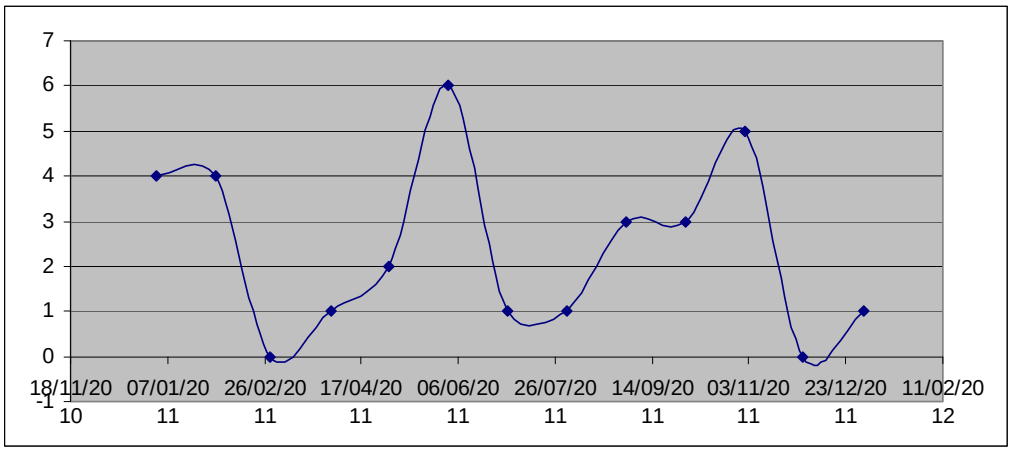


Chart 6 - University of Pisa IR Submission rate without start-up

The Kiviat chart represents a comparison between the MQC quality profile and the profile derived from the CRUI guidelines. The results show that the MQC profile consider this IR better respect to the CRUI profile.

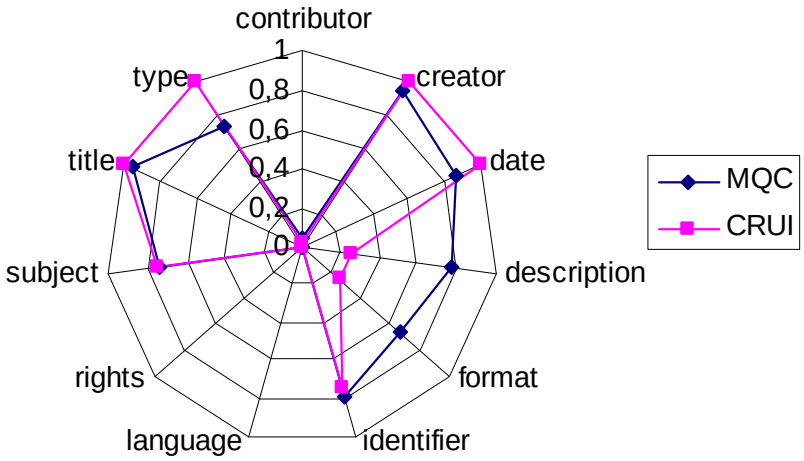


Chart 7

Quality score

University of Bologna

AMS Acta is the University of Bologna's institutional open access repository which collects and disseminates the research output of those operating at the Alma Mater Studiorum, University of Bologna, or taking part in initiatives promoted by its structures.

IR base url: http://amsacta.cib.unibo.it/cgi/oai2

Number of records: 2524
Last harvesting: 2011-12-01
Quality Score: 71653,81

Completeness	
Average	0,749462758
Standard Deviation	0,032071261
Variance	0,001028566
Minimum	0,556
Maximum	0,84
Level of confidence (95,0%)	0,00125178

Accuracy	
Average	0,427
Standard Deviation	0,098749927
Variance	0,009751548
Minimum	0,306
Maximum	0,675
Level of confidence (95,0%)	0,003854327

The chart of the level of Completeness for each field in the IR is very similar to the previous one. The results shows in general that all fields required by the repository system seem filled.
Thus the consideration done for the IR of University of Pisa remain valid here.

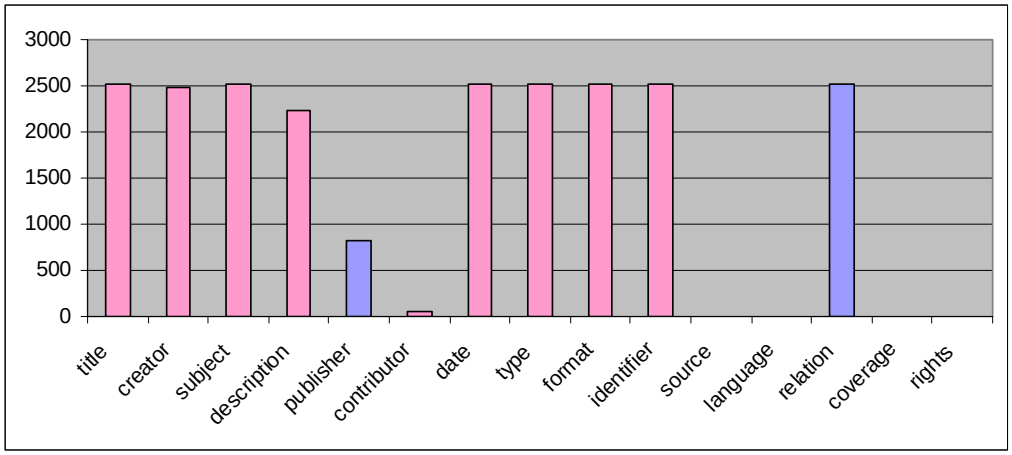


Chart 8 - Fields Completeness for IR of University of Bologna

The accuracy results for this IR is very similar to the previous but the accuracu level of the field Type is lower. This is due to the use of the field. As example, the use of a description like *<DC:Type>Documento relativo ad un convegno o altro*

evento</DC:Type> instead of the CRUI or DRIVER taxonomies, makes inaccurate the field.

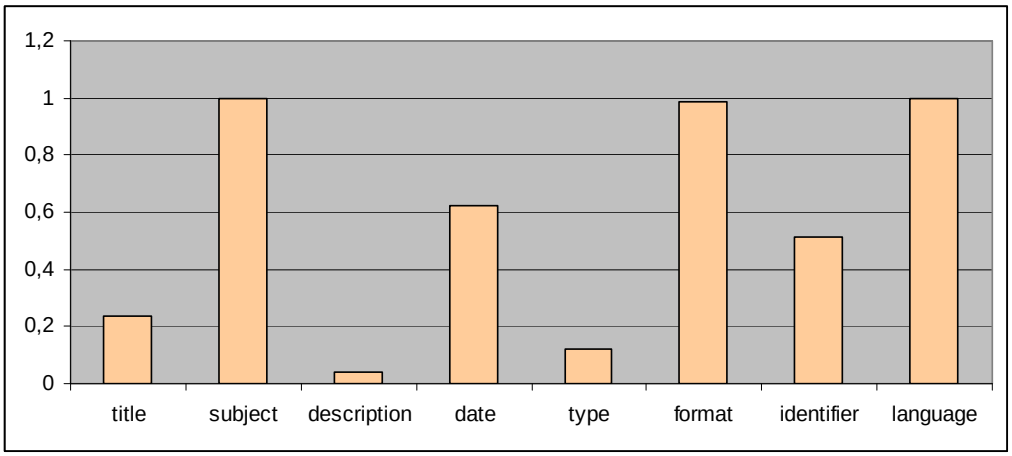


Chart 9 Fields Accuracy for IR of University of Bologna

This chart shows that there is substantial steadiness of the Completeness and the Accuracy level during the time. The Average of the first is stable over the of 0,7 while the latter is stable between the 0,4 and 0,5.

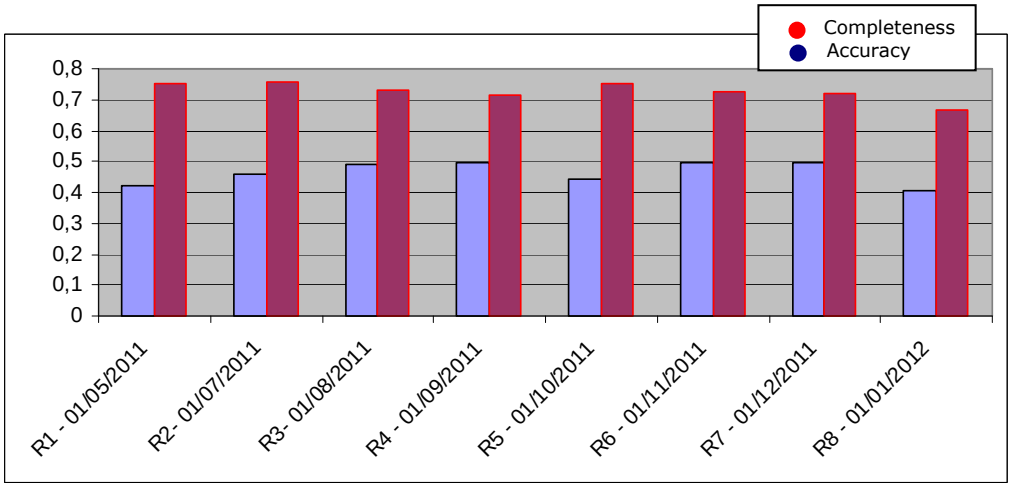


Chart 10 - Time chart of Accuracy and Completeness for IR of University of Bologna

This chart shows a start up action with more than 2000 submissions. After the start up, the IR is maintaining a good "vitality" (chart 11) because the Average of the submission rate after April 2011 is 19/Month submissions.

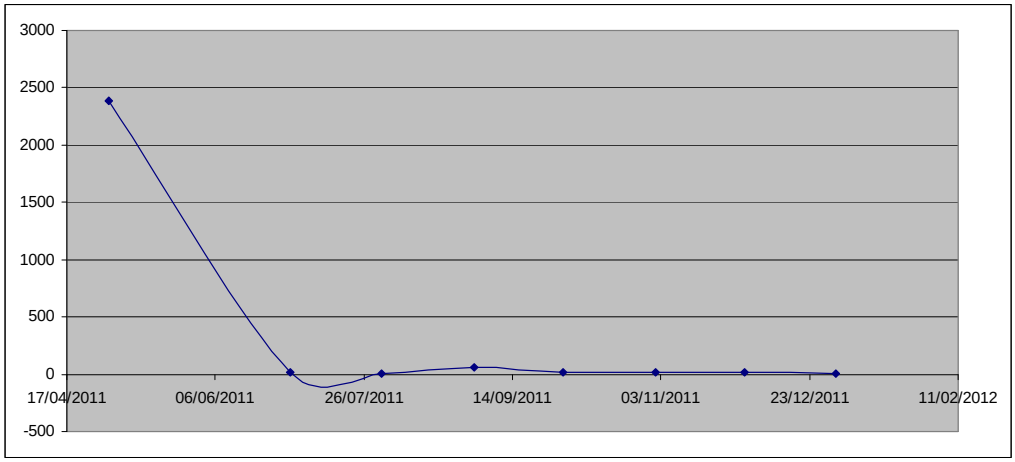


Chart 11 - IR Submission rate - University of Bologna

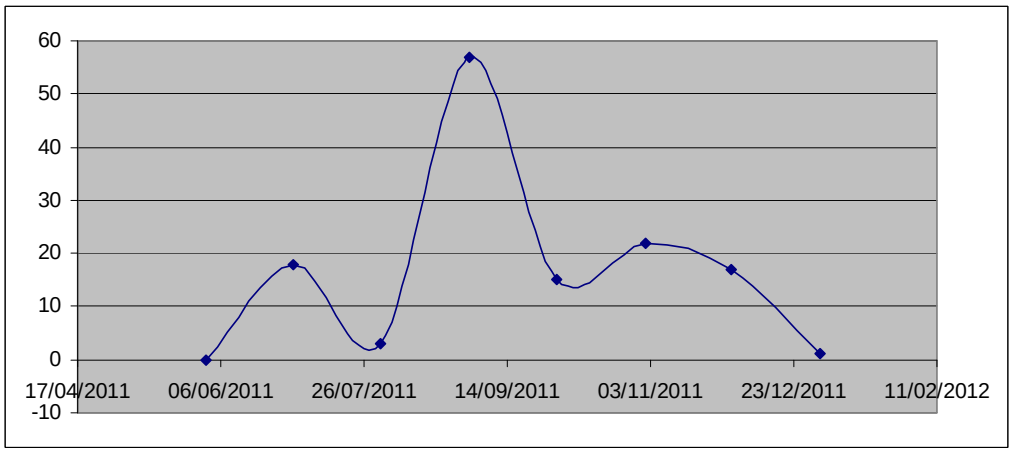


Chart 12 - IR Submission rate without start-up - University of Bologna

Similarly with the IR of Pisa, the MQC profile cover better the effective Completeness of the fields in the IR.

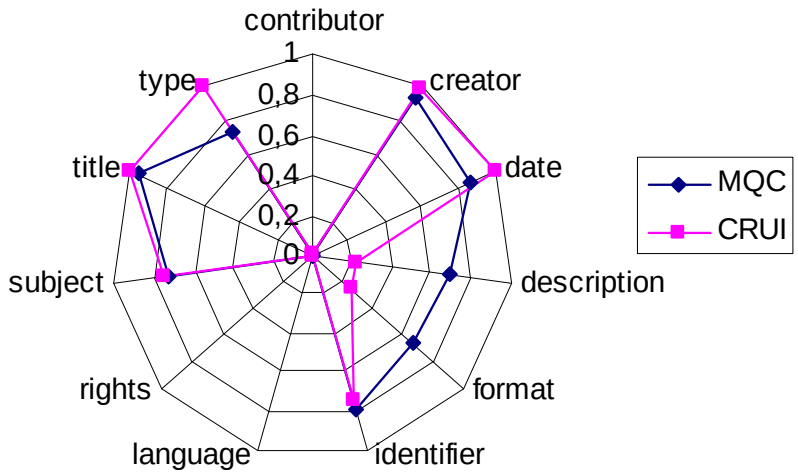


Chart 13 MQC and CRUI quality profiles comparison

ArcA diA – (Archivio Aperto di Ateneo) - University of Roma3

The open IR of the University Roma3 aims to collect and give access to the scientific output of the University, to give effect to the principles of the Berlin Declaration and the Declaration of Messina. The first phase of the project, managed by the University Library System in collaboration with the Office of Research, provides for the publication of the doctoral thesis of the twentieth cycle of doctoral training, discussed in the academic year 2007/2008.

Number of records: 559
Last harvesting: 2012-01-17
Quality Score: 47386,5

Completeness	
Average	0,794275492
Standard Deviation	0,0460801
Variance	0,002123376
Minimum	0,577
Maximum	0,842
Level of confidence (95,0%)	0,003828235

Accuracy	
Average	0,390515
Standard Deviation	0,083399
Variance	0,006955
Minimum	0,28
Maximum	0,607
Level of confidence (95,0%)	0,006929

This IR present a good level of completeness and a reliable workflow. In fact when the field are managed, they have an high filling level while the others are not inserted at all (format, source, coverage).

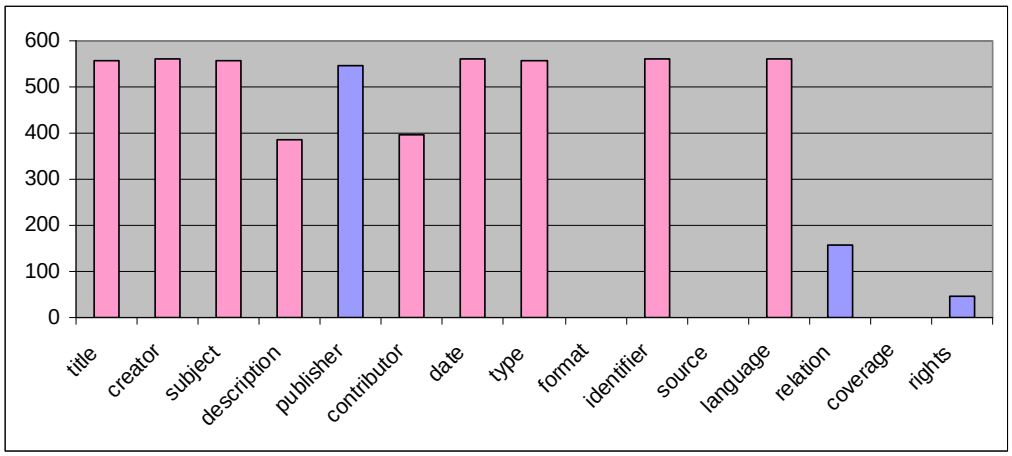


Chart 14 Fields Completeness for IR of University of Roma3

The low level of Accuracy for the field Type is mainly due to an empty space inserted in the field *<DC:Type> Doctoral Thesis<DC:Type>* instead a unique string *DoctoralThesis* ad defined by DRIVER guidelines.

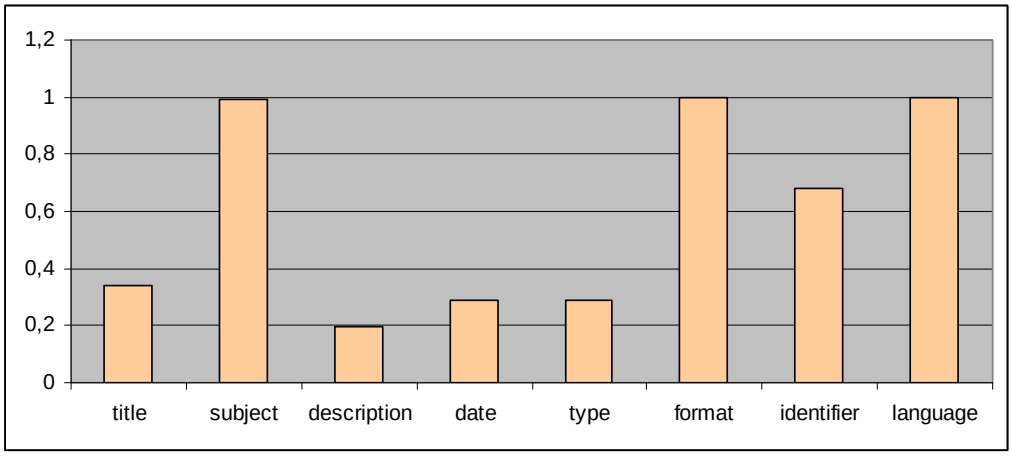


Chart 15- Fields Accuracy for IR University of Roma3

The Completeness line shows that a very good workflow is in place since the substantial steadiness of the high filling score. The Average of the Accuracy level is around 0,5 and follows the same steadiness of the Completeness.

This result can suggest the presence of a systematic issue. In fact, the low but stable level of Accuracy might be caused by a different field implementation rules, for examples the adoption of a different format or encoding, respect to those expected by the MQC. In this case, a specific evaluation has to be done to decide if the new implementation rule can be included in the MQC measurement

criteria in order to extend the range in which a metadata field is considered accurate.

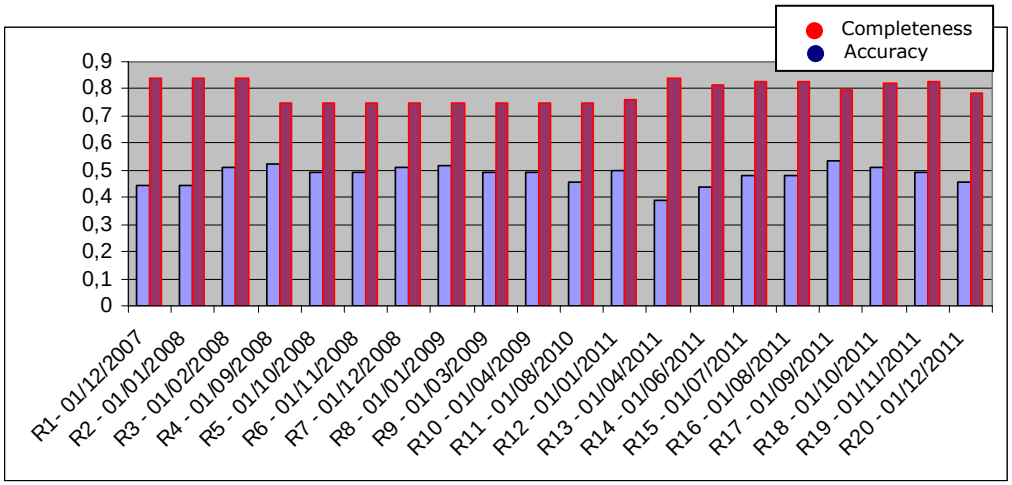
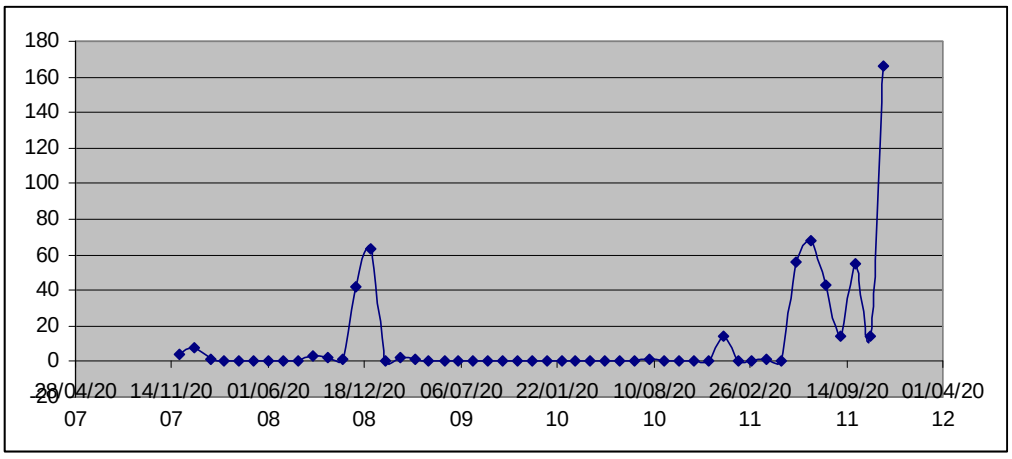


Chart 16 - Time chart of Accuracy and Completeness for IR of University of Roma3

In the submission rate chart, after 3 yeas of substantial inactivity, we assist to an important restart in using the IR with a high peak at the end of the 2011. Thus, if we consider only the 2011, the Average submission rate is over 35/Month submissions. This means a very good "vitality" index.



Char 17 - IR Submission rate University of Roma3

In this case the Kiviat chart seems consider the IR better then the MQC because it consider the filling of some field strongly relevant (Type, Language) respect to MQC. In this case, since fields like Description and Contributor have a low level of Completeness, their impact on the field evaluation is less in the CRUI model respect to MQC.

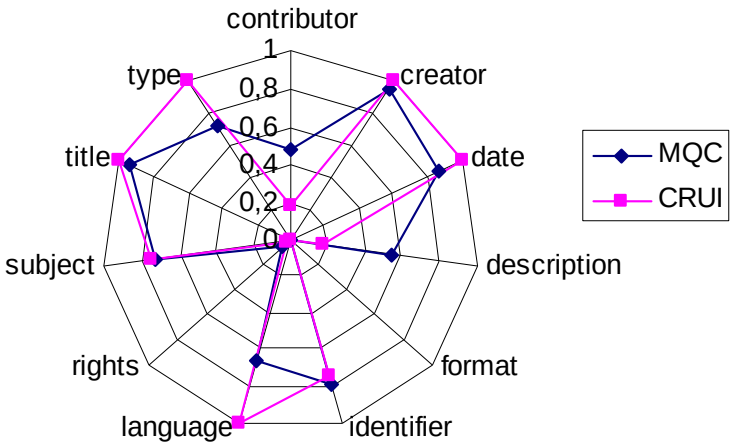


Chart 18 - MQC and CRUI quality profiles comparison

AperTO - University of Turin

This is the OA IR of university of Turin. The IR holds publications, funding research outcomes and doctoral thesis.
IR base url: http://dspace-unito.cilea.it/dspace-oai/request

Number of records:497
Last harvesting: 2012-01-17
Quality Score: 16522

Completeness	
Average	0,816285714
Standard Deviation	0,075339305
Variance	0,005676011
Minimum	0,497
Maximum	0,918
Level of confidence (95,0%)	0,006639762

Accuracy	
Average	0,37159
Standard Deviation	0,080576
Variance	0,006492
Minimum	0,224
Maximum	0,657
Level of confidence (95,0%)	0,007101

The Chart shows an high level of Completeness of the field. In fact all the key fields are well managed with the unique exception of the field Contributor.

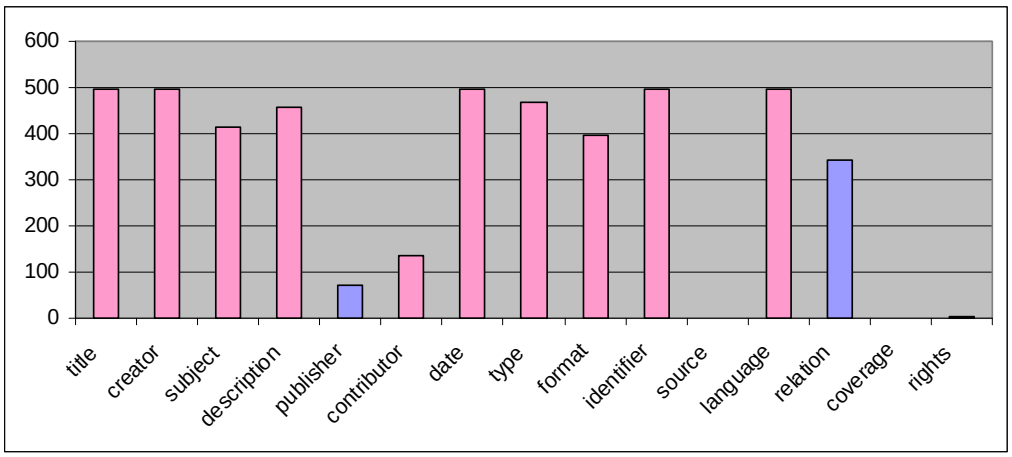


Chart 19 - Fields Completeness for IR of University of Turin

The Accuracy chart, instead, shows some issues. Apart the case of a different encoding/ format of the information, in this case there are some problems related to field Format. This field is considered repeatable and in this IR the second Format instance present a value like this: *<DC:Format> 436292 bytes</DC/Format> that* does not match with the MQC measument criteria. Thus, this value affects the overall Accuracy evaluation of the Format field proportionally.
The Type field presents some not codified values like: "Presentazione" or "Materiale per lezione" tha are not included in the CRUI, DRIVER or MiUR taxonomies.

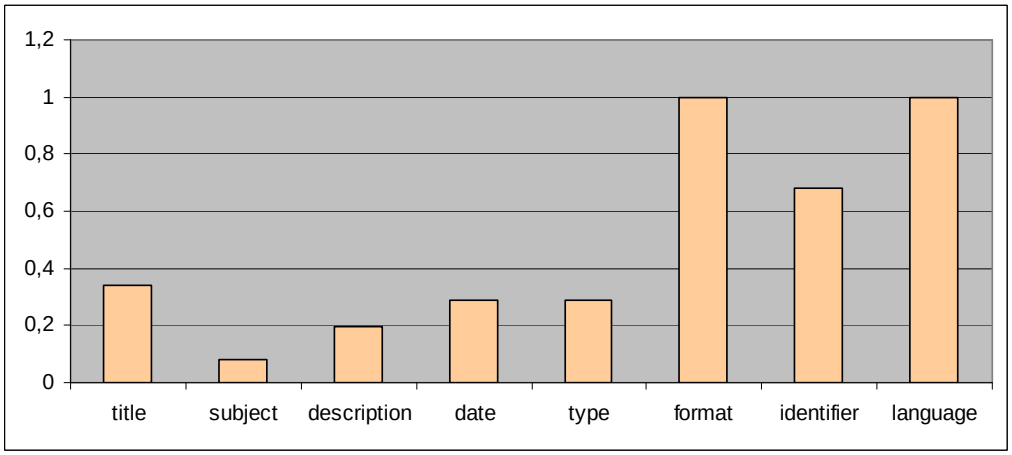


Chart 20 - Fields Accuracy for IR University of Turin

The Chart 21 presents an interesting inverse correlation between Completeness and Accuracy (-0,6827) that require a further

analysis. In fact could be interesting understand which is the critical field that every time it is used, it provokes the Accuracy loss.
According to the Chart 21 and Chart 20 the critical fields cold be the Subject and Type. In fact they are not always filled and at the same time have a very low level of Accuracy. Thus we might assume that every time the Subject and/or Type field are filled, the Accuracy score falls down because of the wrong values inserted.

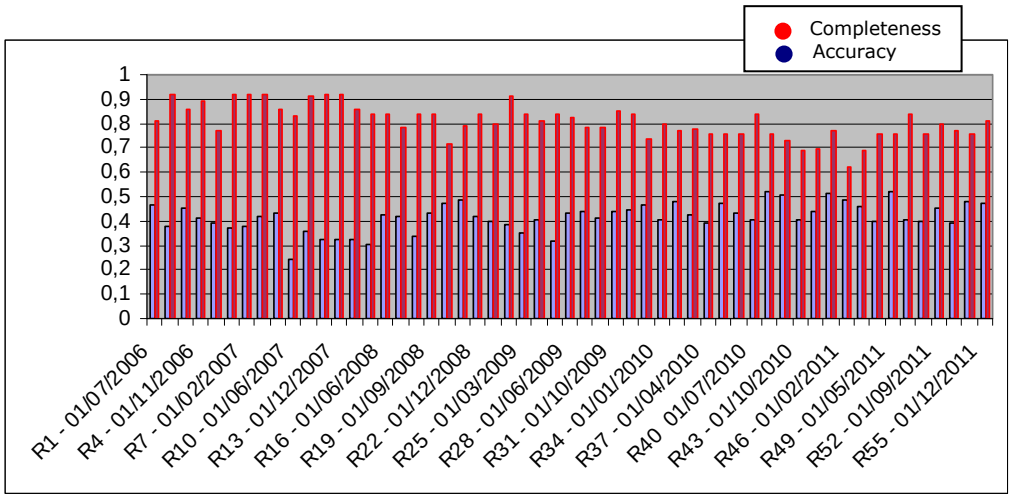


Chart 21 - Time chart of Accuracy and Completeness for IR of University of Turin

This is an "old" IR since the first publications were submitted in the 2006.
The chart shows a sort of activity concentrated in two moments (peaks) in the past while now the submission rate is very low.

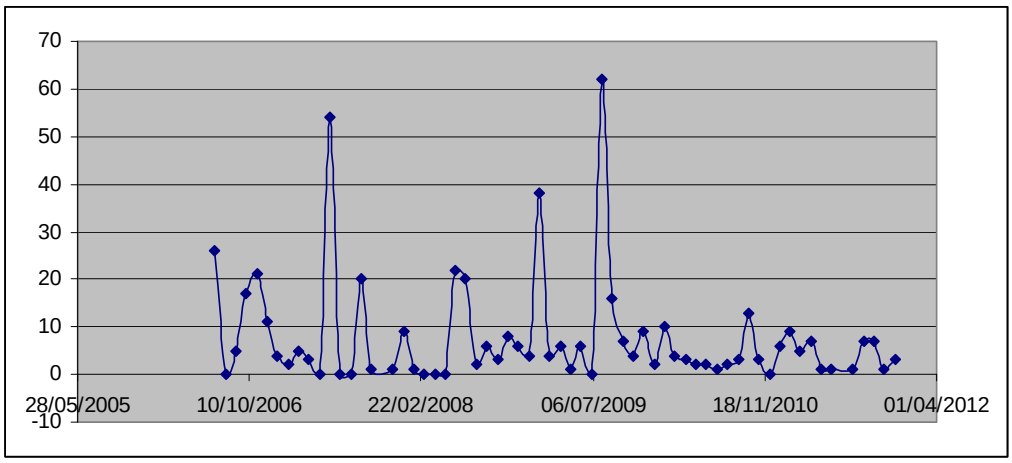


Chart 22 - IR Submission rate University of Roma3

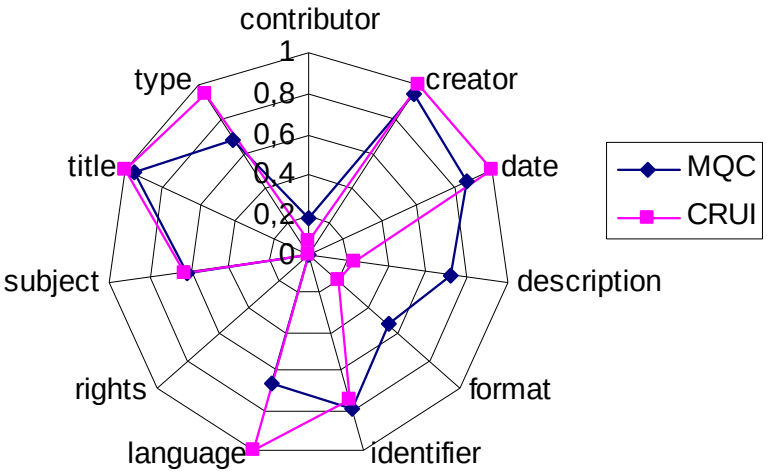


Chart 23 - MQC and CRUI quality profiles comparison

University TorVergata

Ir base URL: http://art.torvergata.it/dspace-oai/request
Number of records:14866
Last harvesting: 2012-01-17
Quality Score: 26957,12

Completeness	
Average	0,786491053
Standard Deviation	0,048640988
Variance	0,002365946
Minimum	0,552
Maximum	0,922
Level of confidence (95,0%)	0,000781968

Accuracy	
Average	0,501088
Standard Deviation	0,107573
Variance	0,011572
Minimum	0,249
Maximum	0,72
Level of confidence (95,0%)	0,001729

This is the most populated repository analyzed in this research. Despite the high number of publication managed, the overall level of Completeness is high. As we said, this suggest that a reliable submission workflow is implemented in the institution.

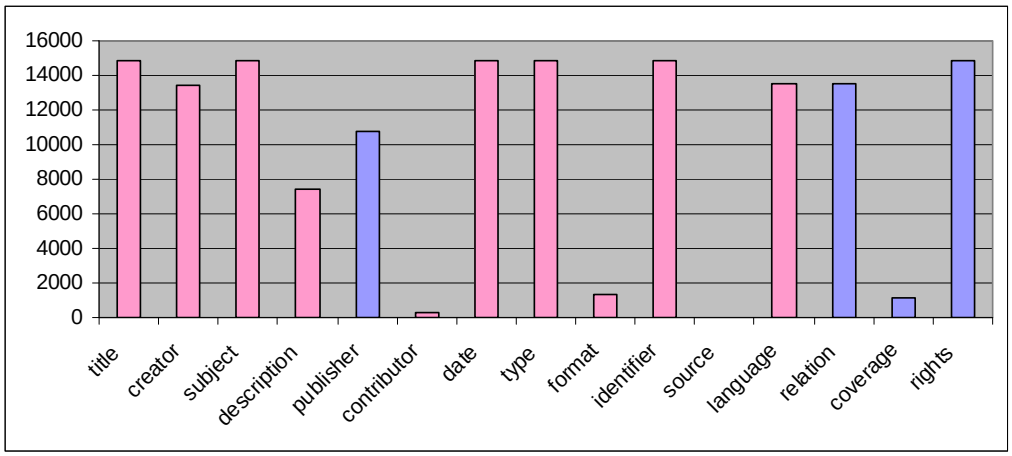


Chart 24 - Fields Completeness for IR of University of TorVergata

Unfortunately, the level of Accuracy dens not follow the Accuracy trend. The Subject issues are related to the out-of standard values like <DC:Subject> Ricerca Cardiovascolare ed Ematologica</DC:subject>

The identifier field is quite accurate but it is worth to notice the following case.

This is a value detected in the identifier fied that the MQC measurement has considered inaccurate <DC:identifier> 10.1016/j.cardiores.2004.07.024</DC:identifier>. At first glace this seems a DOI identifiers but the missing fo the namespace or the resolver URL makes it indecipherable thus, unusable.

A value detected in the date field and reported here as example is: <DC:date> 2008-04</DC:date>. In this case it is impossible to interpret correctly the value of "04".

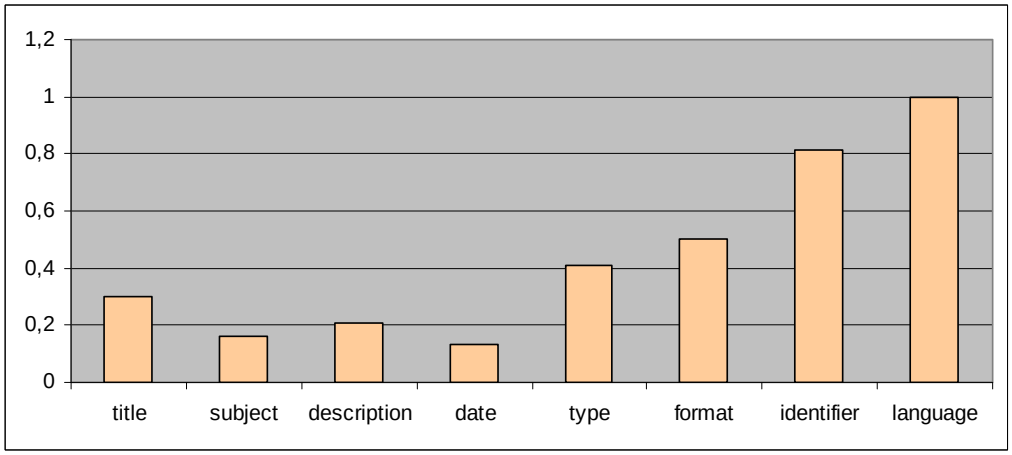


Chart 25 - Fields Accuracy for IR University of TorVergata

This IR was populated through 3 massive submission (Chart 27). In order to maintain the stable level of Completeness and Accuracy it is possible that metadata associated to these objects are generated in the same way and at the same time.
Moreover, it is worth to notice that a massive submission can be performed by the institutional staff only.

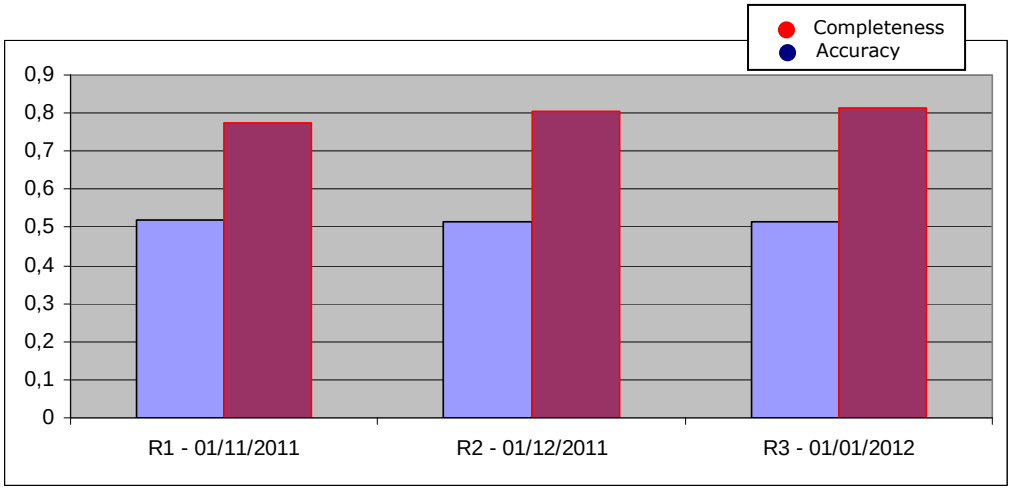


Chart 26 - Time chart of Accuracy and Completeness for IR of University of TorVergata

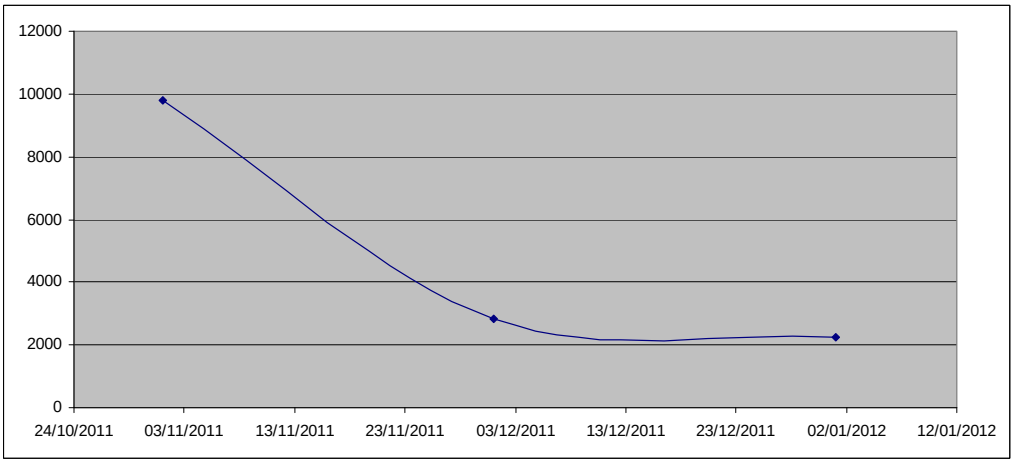


Chart 27 - IR Submission rate University of TorVergata

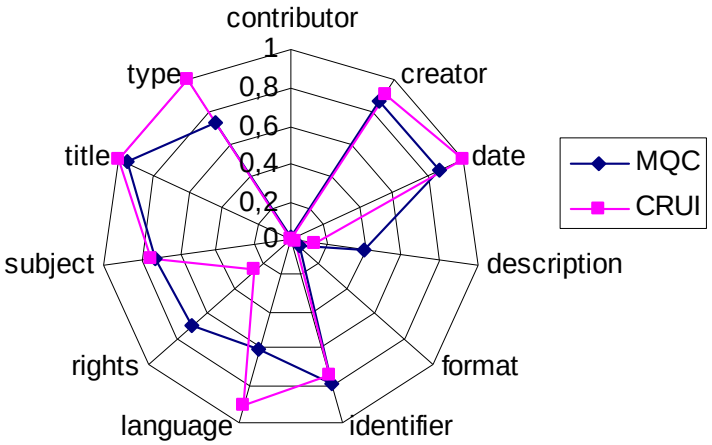


Chart 28 MQC and CRUI quality profiles comparison

University of Parma

IR base URL: http://dspace-unipr.cilea.it/dspace-oai/request
Number of records: 1128
Last harvesting: 2012-01-17
Quality Score: 26957,12

Completeness	
Average	0,860664894
Standard Deviation	0,076839127
Variance	0,005904251
Minimum	0,555
Maximum	1
Level of confidence (95,0%)	0,004488928

Accuracy	
Average	0,432593
Standard Deviation	0,070872
Variance	0,005023
Minimum	0,28
Maximum	0,698
Level of confidence (95,0%)	0,00414

The Chart 29 shows a very good level of completeness. In fact, the average is 0,86. This IR manages also those fields less use such as relation or publisher.

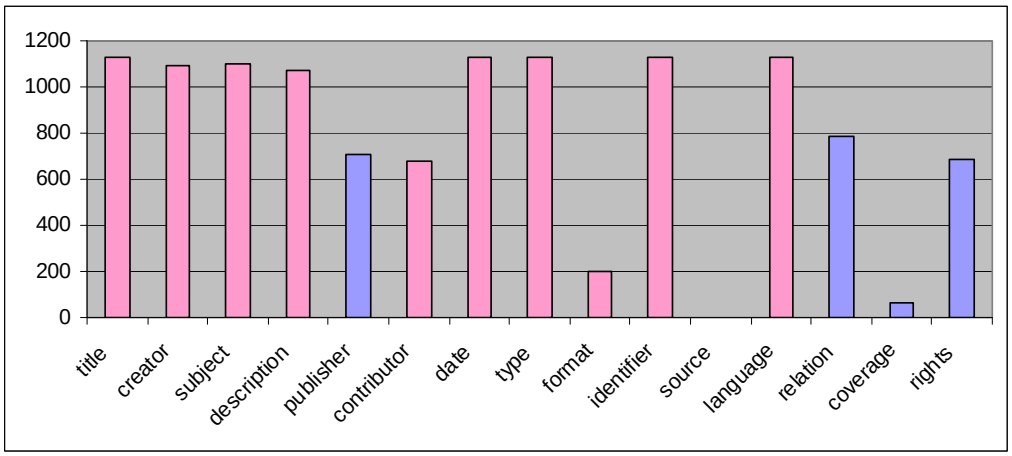


Chart 29 - Fields Completeness for IR of University of Parma

Unfortunately to the high level of completeness does not correspond the same level of accuracy as shown by the Chart 30.

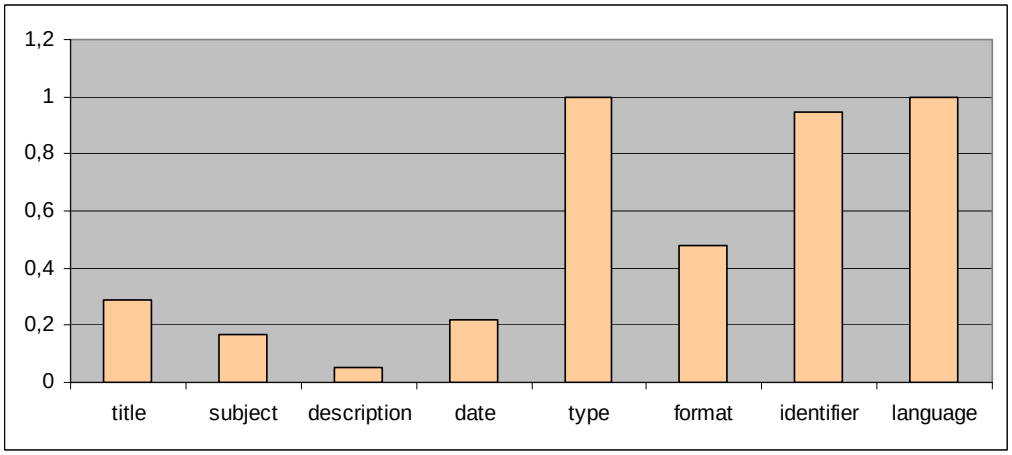


Chart 30 - Fields Accuracy for IR University of Parma

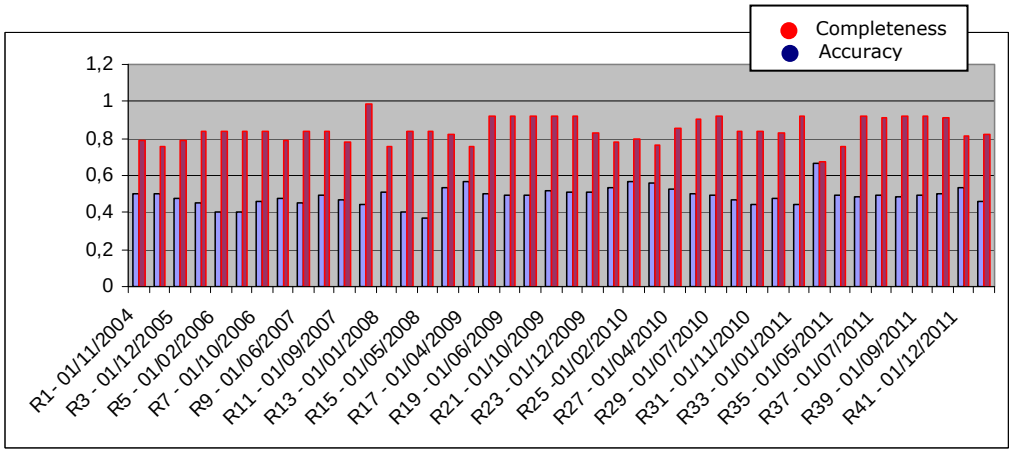


Chart 31 - Chart of Accuracy and Completeness for IR of University of Parma

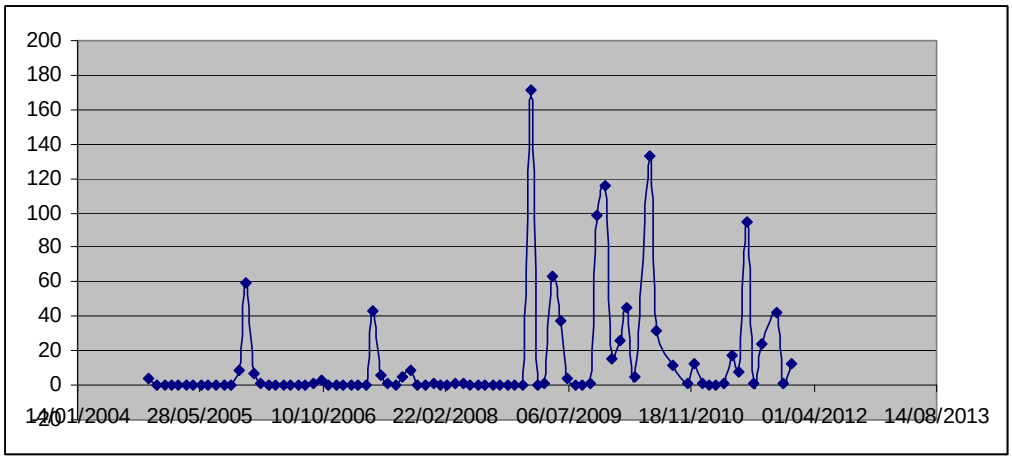


Chart 32 - IR Submission rate University of Parma

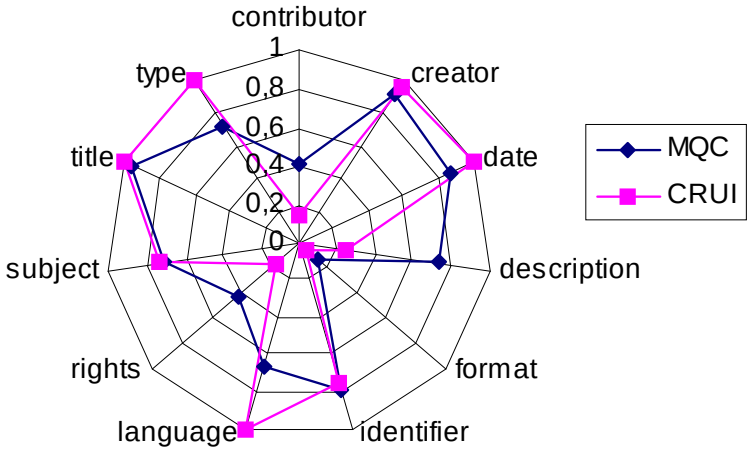


Chart 33 - MQC and CRUI quality profiles comparison

University of Trieste

IR base URL: http://www.openstarts.units.it/dspace-oai/request
Number of records: 5027
Last harvesting: 2012-01-17
Quality Score: 4301,673

Completeness	
Average	0,709204297
Standard Deviation	0,114513524
Variance	0,013113347

Minimum	0,273
Maximum	1
Level of confidence (95,0%)	0,003166322

Accuracy	
Average	0,436568
Standard Deviation	0,099032
Variance	0,009807
Minimum	0,201
Maximum	0,698
Level of confidence (95,0%)	0,002738

The Chart 34 shows a variability in the field filling. The fields as Title, Date, Identifier and Language are fully complete while for the others there is an high level of unpredictability.

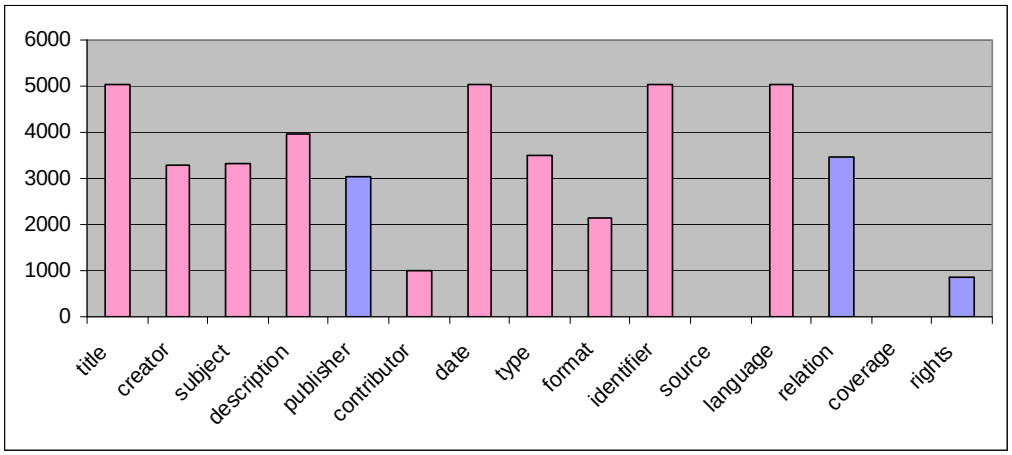


Chart 34 - Fields Completeness for IR of University of Trieste

The Accuracy presents some issues in the subject field the problem is the use of out-of–standard values like: <DC:Subject> prospettive di sviluppo dei traffici nell'Adriatico</DC:Subject>. The same issue was detected for the Format field with the number of bytes and strange values like that:<DC:Format> 5 14"</DC:Format>

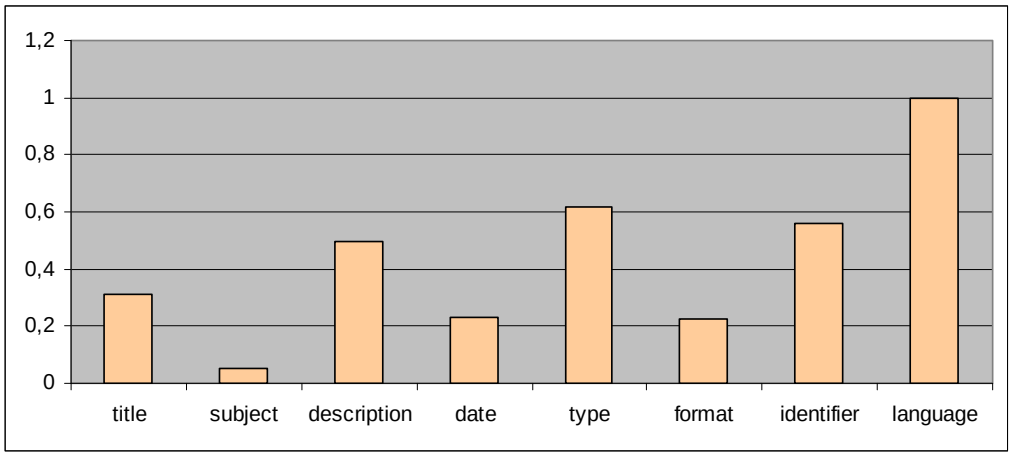


Chart 35 - Fields Accuracy for IR University of Trieste

The Chart 36 shows that the variability has always been there over time. This may be the result of too flexible workflow and/ or unclear policies and guidelines.

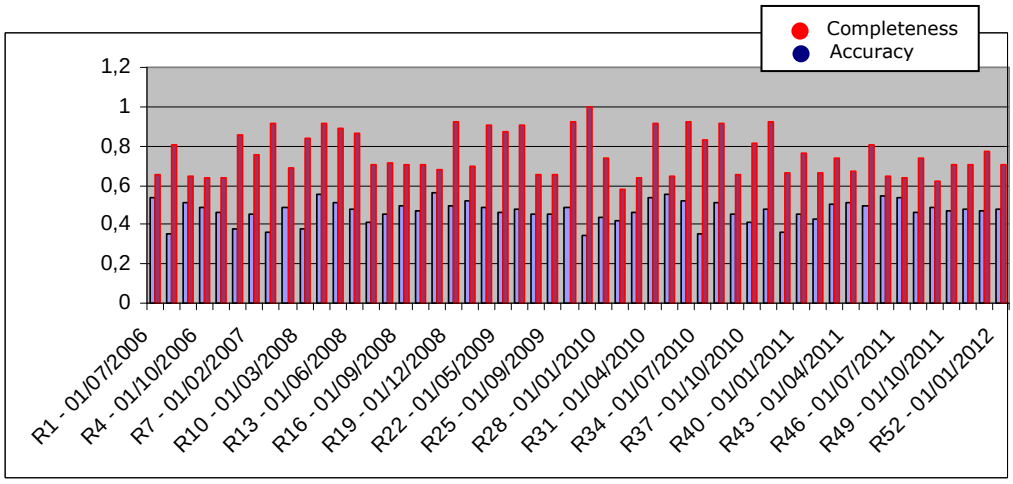


Chart 36 - Chart of Accuracy and Completeness for IR of University of Trieste

Despite this variability in the completeness and accuracy, the IR has a very good "vitality" with a submission rate over the 190 contents/Month in the last year (2011).

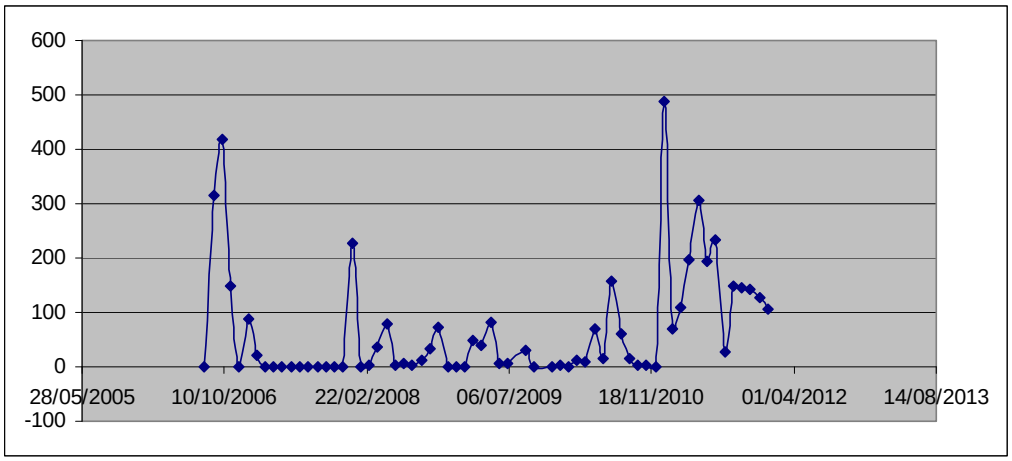


Chart 37 - IR Submission rate University of Trieste

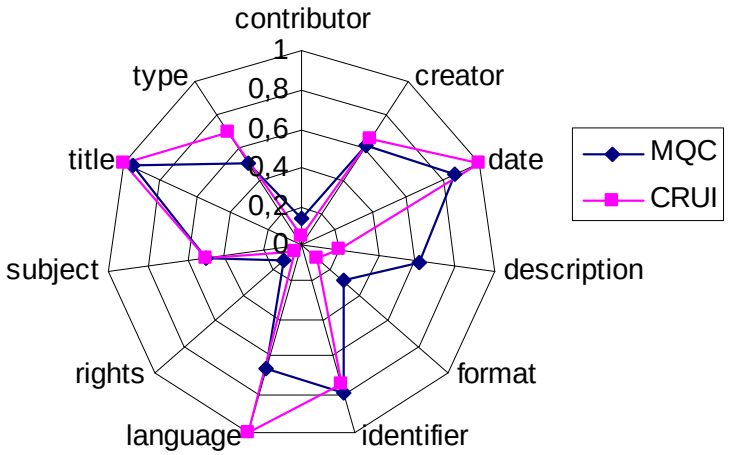


Chart 38 - MQC and CRUI quality profiles comparison

University of Trento

IR base URL: http://eprints.biblio.unitn.it/perl/oai2
Number of records:1587
Last harvesting: 2012-01-17
Quality Score: 103737,5

Completeness	
Average	0,750543163
Standard Deviation	0,032593795
Variance	0,001062355
Minimum	0,47
Maximum	0,76

Level of confidence (95,0%)	0,001604819

Accuracy	
Average	0,38012
Standard Deviation	0,079543
Variance	0,006327
Minimum	0,29
Maximum	0,598
Level of confidence (95,0%)	0,003916

The chart 39 shows an impressive results of the field completeness. The total missing of values for the field Language, Relation or Rights might be due to the impossibility of inserting of a value through the user interface or the strict respect of an institutional policy or a crosswalk issue that prevent the exposing of some fields through the OAI-PMH protocol.

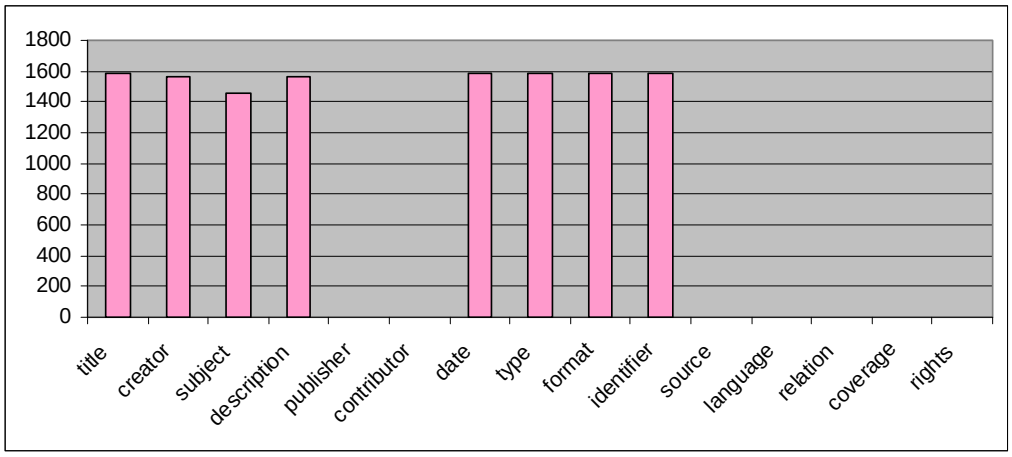


Chart 39 - Fields Completeness for IR of University of Trento

The accuracy results are very interesting. In fact, for the subject field are used out-of-standard values like <DC:Subject>HD Industries. Land use. Labor</DC:Subject>.
The same situation is valid for the Type field. An example of the out-of–standard value is <DC:Type>Departmental Technical Report/<DC:Type>
For the field Format we detected a deprecated use of the field.
In fact, a common behaviour was put together the extension of the file format and the URL to the resource. For instance:
<DC:Format>pdf
http://eprints.biblio.unitn.it/archive/00000014/01/1_99_leonardi.pdf
</DC:Format>

This issues can be solved harvesting the metadata in the MPEG21 or METS format.

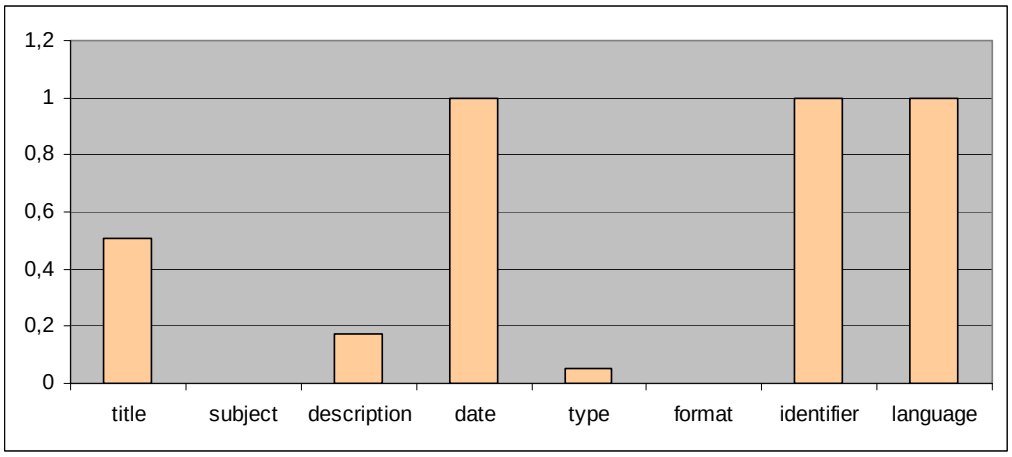


Chart 40- Fields Accuracy for IR University of Trento

The Chart 41 suggest that the a stable workflow is in place. In fact, the Average of the completeness is stable since the first submission. The accuracy instead has some oscillation and the value range from 0,4 to 0,5.

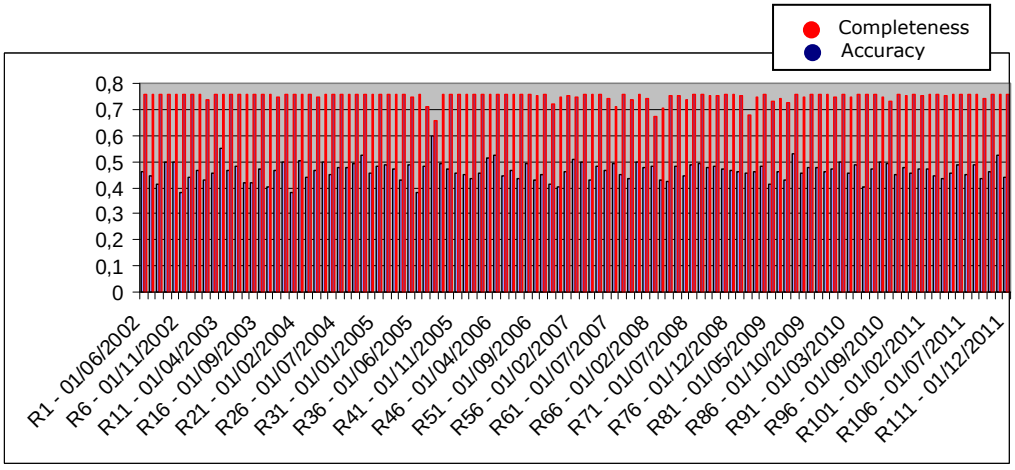


Chart 41 - Chart of Accuracy and Completeness for IR of University of Trento

The Cart 42 shows a low but constant activity around the IR during the time. In the last two years this activity is becoming more important with 3 peaks that have increased the number of publications inside the IR significantly.

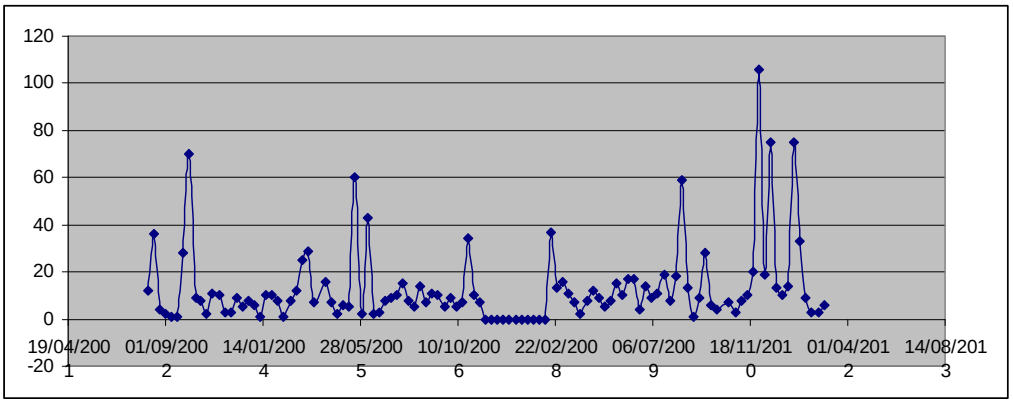


Chart 42 - IR Submission rate University of Trento

In the Char 43 is evident the differences between the MQC QP and CRUI derived QP. In fact, even if for the Creator and Tile the evaluation is very similar, a important difference comes out when are evaluated the Format and Description field.

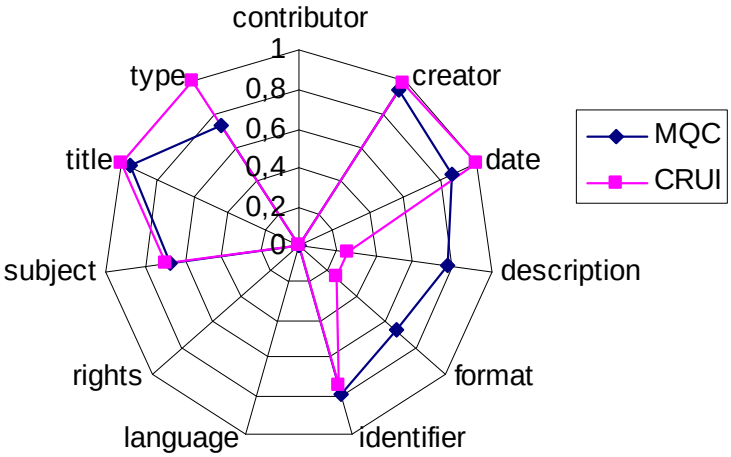


Chart 43 - MQC and CRUI quality profiles comparison

University of Tuscia

IR base URL: http://dspace.unitus.it/dspace-oai/request
Number of records:2093
Last harvesting: 2012-01-17
Quality Score: 10917,51

Completeness	
Average	0,780772575
Standard Deviation	0,080508446
Variance	0,00648161
Minimum	0,488
Maximum	1
Level of confidence (95,0%)	0,00345109

Accuracy	
Average	0,371672
Standard Deviation	0,092406
Variance	0,008539
Minimum	0,174
Maximum	0,698
Level of confidence (95,0%)	0,003961

Similarly to other IR, the field effectively used are mostly completed with the unique exception of the field Rights that is not included in our quality assessment.

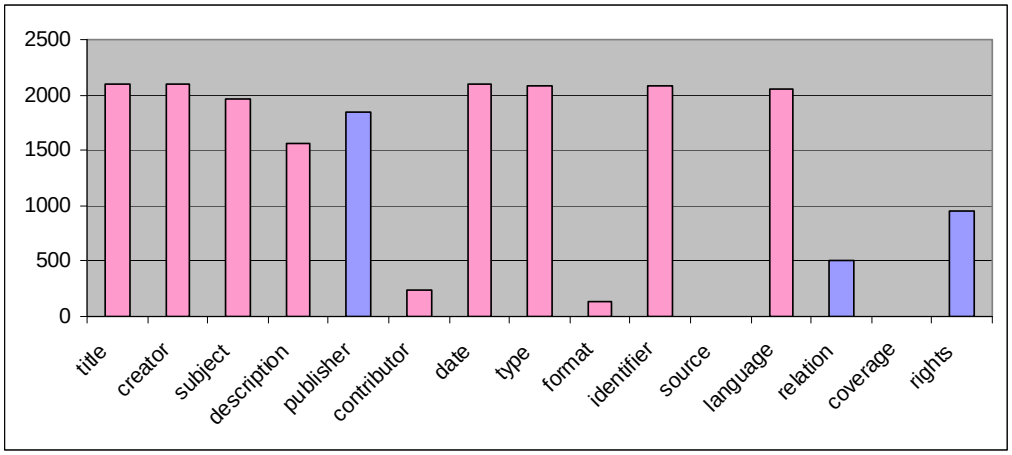


Chart 44 - Fields Completeness for IR of University of Tuscia

The Chart 45 shows a very low level of Accuracy for Subject, Format and Identifiers fields. For the Subject field the main problems is its wrong use. As example we reported some values detected in the IR:
<DC:subject>PDF</DC:Subject>
<DC:subject>RSS feeds</DC:Subject>
<DC:subject>Topic map</DC:Subject>
<DC:subject>Cyberpunk</DC:Subject>
<DC:subject>Domenico Grimaldi</DC:Subject>

Here the use of the Subject field is similar to the use of keywords. This might happen if there is a crosswalk software problem or a misunderstand of the field.
In the field Format we detected number of bytes (e.g. 66924 bytes)
In the Identifier field were detected values as the following:
<DC:identifier>1-4020-1631-X<DC:identifier>
<DC:identifier>F. SAGGINI, Women in British Romantic Theatre and Drama in La questione Romantica, nr. 9. Liguori editore, Napoli, 2000. pp. 234-241.<DC:identifier>

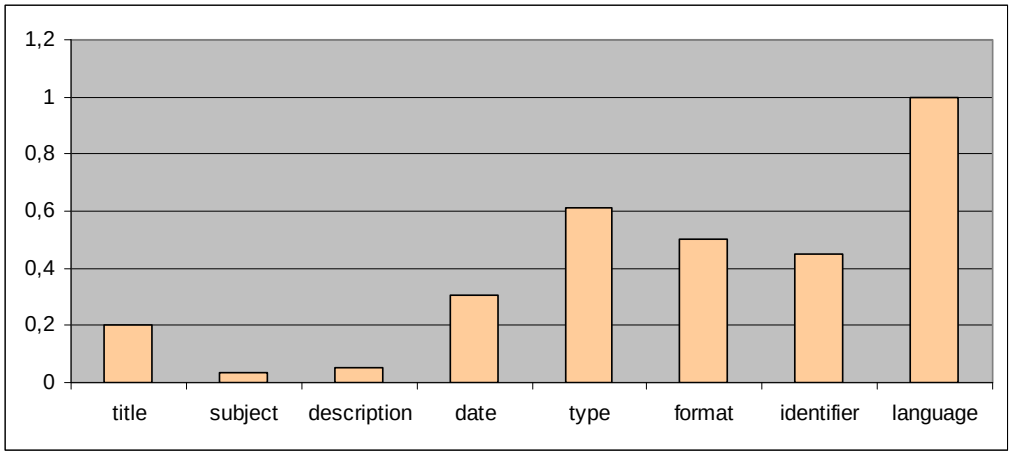


Chart 45 - Fields Accuracy for IR University of Tuscia

The Chart 46 shows some oscillations in the Completeness but the values are stable over 0,7.
The accuracy instead, present several oscillations. Sometimes an high level of accuracy is related to a low level of completeness.
This can be due to several factors and a deeper analysis is needed.

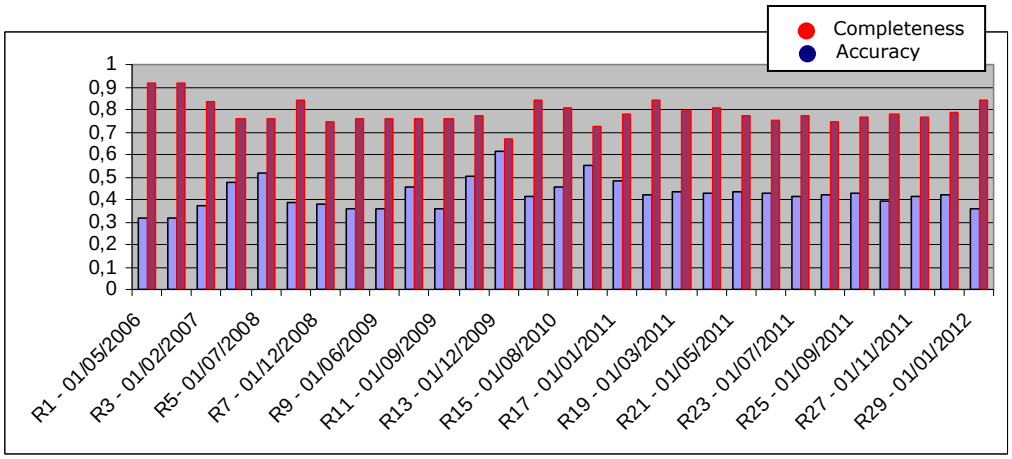


Chart 46 - Chart of Accuracy and Completeness for IR of University of Tuscia

The Chart 47 shows that after 4 years of substantial inactivity the IR activity is strongly restarted. The distribution shows 3 main peaks that can be referred to the thesis submissions.

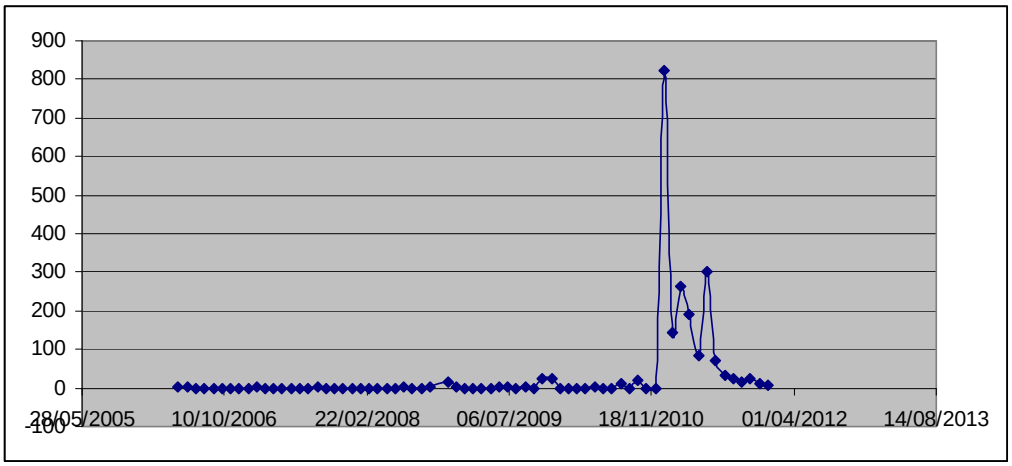


Chart 47 IR Submission rate University of Tuscia

The Chart 48 shows how the MQC QP covers better the effective level of IR completeness. In fact the MQC more rewards field mostly complete as Description respect to CRUI derived QP.

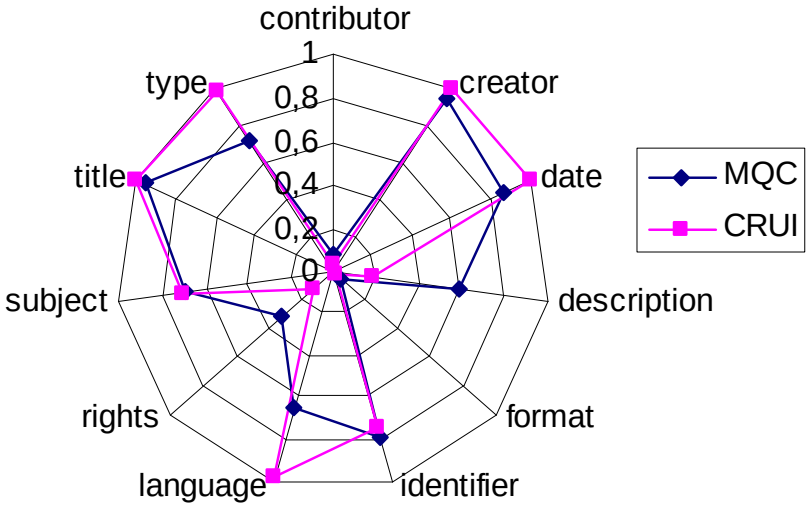


Chart 48 - MQC and CRUI quality profiles comparison

Chapter 8

Conclusions

8.1 Metadata Quality Assessment Results

The OA movement is growing more and more in the academic world and it is expected that future executive's model and new forms of dissemination of knowledge are supported by this tool.
The effectiveness of the OA instrument is mainly based on both the quantity and club quality of the research results made freely available, but it is not enough. In fact, to be effective in the OA resource reuse and exploitation it 'requires that these contents are searchable and available on the internet.
To tackle this challenge, this research defines and tests a Quality Framework for descriptive metadata of the IR. From the results we can stated that:

a) The Completeness seems to be well addressed by all IR analyzed. Moreover we noticed that there are few cases in which the fields are used, let say, randomly (with a probability of 50%). In fact, statistics show that either the fields are filled in all the IR records, or not at all. This suggests that workflows are stabilized on the use of certain fields than others, but those selected are well managed.

b) There are some issues in the Accuracy dimension. The major problems were detected on the free – text fields such Title and Description. This is not a surprise given that the free text fields are suitable for flexible use. Unfortunately, this flexibility results in an uncontrolled use of the fields causes accuracy issues. For instance the use of description in more then one language in the same description field is considered a bed practice. Instead there was a surprise on the subject field. In fact, the existence of the subject headings of the MiUR should suggest that the use of an authority file or a predetermined values list for that field. A possible explanation

of this situation can be given considering the difficulty and the scarcity of resources dedicated, up to now, to OA IR implementation. This may have caused a slowdown in maintenance and upgrades.

c) The DC is not expressive enough to support the complexity of the resources and their descriptive needs. In fact, the dumb down process forces values linked to the principal field into the same field. This is an Accuracy problem because the value expected is different to the value detected. For instance, the MIME format of the resource is assigned to the Format field but also other related information such as the number of bytes of the file, a concatenation of the MIMA and the link to the resource itself , and so forth, fall into different instances of the same field.

d) We observed in two cases an interesting inverse correlation between the Completeness and the Accuracy that should require a deeper analysis.

e) We showed the validity of MQC model respect to those derived from the CRUI guidelines. The translation of the CRUI guidelines into a QP is arbitrary and was useful at high level to analyze the presence of macro differences.

f) There are some cases in which the values could be considered accurate but their encoding format was not included in the our measurement model. These cases partially contribute to a low score of Accuracy. It is evident that a community consensus should be sought not only to define the QP, but also to determine shared measuring modalities.

8.2 Possible improvement actions

The improvement actions should be prioritized is based on the severity of quality issues detected according to the metrics defined.
In [Bruce, Hilmann, 2004] are described a sort of model in three tiers with a number of actions associated that can be used as a

inspiration to define the priority actions that an institution has to perform to solve the quality issues. These actions have to take into account their impact on the archive quality. For example, if there are some records where the titles are missed, actions to solve this problem have more impact on the metadata quality of the repository respect actions oriented to face issues on DC:rights field.
To this end the service outcome should not be only the list of the problems detected in the repository, but a report of actions ranked on the base of the criticism.

1) First tier of actions (completeness)
The first tier of action is related to the elements with an high weigh associated that are empty or incorrect.

To fill (with controlled vocabularies if required) the elements that are empty starting from those that are mandatory in the guidelines and have a strong impact on the metadata function requirement such as discoverability and accessibility evaluate with the field usage statistics.

2) Second tier of actions (accuracy)

- to fill/correct (with controlled vocabularies if required) fields with a strong impact on the overall metadata quality estimation. Examples cold be the definition of formats of the digital resources or the language definition.

3) Third tier of actions (consistency)
- to fill/correct (with physical check) the fields resulted inconsistent. Examples are DC:identifier and DC:format.

8.2 Next steps

This research set up the condition for further analysis and refinements. For instance, since has been shown that unqualified DC is not expressive enough for a detailed analysis, a new assessment should be performed using MPEG21-DIDL or METS. This analysis is useful in particular from the MQC service point of view because allows an estimation of the costs in term of complexity, time, efforts for obtaining a finer metadata quality evaluation.

Some clues of a possible relation between the accuracy and completeness were found out. Investigating this aspect could be useful , in particular for the Interaction designer and developers that might have new information for design submission interface able to rise up the level of Accuracy of the metadata.

Finally, this analysis is intended to be a stimulus for the institutions that subscribe to OA movement for improving the metadata quality inside the IR in order to make the Knowledge actually Open and Accessible.

Bibliography

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: Supporting communities of discourse and practice - metadata research and applications, Seattle, Washington, USA, 28 September - 2 October 2003, Stuart A. Sutton, Jane Greenberg and Joseph T. Tennis (Eds.), Syracuse (NY): Information Institute of Syracuse, 2003, p. 39-48, http://dcpapers.dublincore.org/ojs/pubs/article/view/732/728

Basili V.R, Caldiera G. and Rombach H. D., "The goal question metric approach", in Encynclopedia of Software Engineering. Wiley, 1994.

Basili V.R, GQM Gold practice https://goldpractice.thedacs.com/practices/gqm/index.php , 2005

Bellini Emanuele, Deussom Marcel Aime, Nesi Paolo, Assessing Open Archive OAI-PMH implementations – DMS2010

Bellini Emanuele, Nesi Paolo - A Trust P2P network for the Access to Open Archive resources. IFLA Conference 2009, Milan

Bellini P., Bruno I., Nesi P., "Visual Programming of Content Processing Grid", The 15th International Conference on Distributed Multimedia Systems, DMS2009.

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

Blake, Miriam E. and Frances L. Knudson. "Metadata and Reference Linking." Library Collections, Acquisitions & Technical Services 26 (3), (2002): 219-230

Botella P. et al, ISO/IEC 9126 in practice: what do we need to know?

Brody Tim, Carr Les, Harnad Stevan, Swan Alma - "Time to Convert to Metrics" Research Fortnight 18 July 2007 pp 17-18. http://users.ecs.soton.ac.uk/harnad/Temp/refortnight.pdf

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)

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 http://idea.library.drexel.edu/handle/1860/1600

Burnett, K., Ng, K., & Park, S. (1999). A Comparison of the Two Traditions of Metadata Development. Journal of the American Society for Information Science, 50(13), 1209-1217.

CARROLL, J. M. (1995). Introduction: the scenario perspective on system development. In J. M. Carroll (Ed.) *Scenario-based design: envisioning work and technology in system development* (pp. 1-18). New York: John Wiley & Sons, Inc

Donatelli A, Longobardi R, Gangemi R, Marinelli C,Unified Scenario-Based Design, Part 1: Methodological principles – IBM - 2005 http://www.ibm.com/developerworks/rational/library/05/1129_donatelli/

Efron Miles: Metadata Use in OAI-Compliant Institutional Repositories. J. Digit. Inf. 8(2): (2007)

EUROHORCs' Recommendations on Open Access, 2008: http://www.eurohorcs.org/SiteCollectionDocuments/EUROHORCs_Recommendations_OpenAccess_200805.pdf

Evans & Lindsay – The management of quality control (6 ed.) Mason, OH Thompson

Fenton, N. E., and Pfleeger, S. L. *Software Metrics – A Rigorous and Practical Approach, 2nd Edition,* PWA Publishing Company, Boston, MA, 1997.

Forneser, Brurteau, Shrum, CMMI for Service CMMI-SVC version 1.2Geneva: International Organization for Standardization

Fouloneau Muriel and Francis André - Investigative Study of Standards for Digital Repositories and Related Services - DRIVER project

Garvin D., "What Does "Product Quality" Really Mean?" *Sloan Management Review*, Fall 1984, pp. 25-45

Greenberg, J. (2001). A Quantitative Categorical Analysis of Metadata Elements in Image-Applicable Metadata Schemas. Journal of the American Society for Information Science, 52(11), 917-924

Gudrun Fischer, Norbert Fuhr- Heterogeneity in Open Archives Metadata- Cyclades project

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

Harnad, S. (2007) The Green Road to Open Access: A Leveraged Transition. In: *The Culture of Periodicals from the Perspective of the Electronic Age*, pp. 99-105, L'Harmattan

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y, Oppenheim, C., Stamerjohanns, H., & Hilf, E. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials Review, 30(4). pp 310-314, ISSN 0098-7913, doi:10.1016/j.serrev.2004.09.013 http://eprints.ecs.soton.ac.uk/15753

IETF RFC 2045 Multipurpose Internet Mail Extensions (MIME), 1996

IFLA - Functional Requirements for Bibliographic Records FRBR, Final Report http://archive.ifla.org/VII/s13/frbr/frbr1.htm. 1998

IFLA Cataloguing Section Working Group on the Use of Metadata Schemas - Guidance on the Structure, Content, and Application of Metadata Records for Digital Resources and Collections, , 2003, http://archive.ifla.org/VII/s13/guide/metaguide03.pdf

ISO 14721:2003 Reference Model for an Open Archival Information System (OAIS)

ISO 25000 SquaRE ISO/IEC 25020 : Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Measurement reference model and guide - 2005 - ISO/IEC JTC1/SC7/WG6; ISO/IEC 25021 Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Quality measure elements

ISO/IEC 9124: Software Engeeneering-Product quality part1,2 and 3 Geneve, 2001 (part 1) 2003 (part 2 and 3)

ISO/IEC IS 15939, Software Engineering – Software Measurement Process, 2002

Jewel Hope Ward, QUANTITATIVE ANALYSIS OF DUBLIN CORE METADATA ELEMENT SET (DCMES) USAGE IN DATA PROVIDERS REGISTERED WITH THE OPEN ARCHIVES INITIATIVE (OAI) dissertation

Jura, J - Juran on quality by design - New York, NY 1992 Free press

Kelly Brian, Closier Amanda, Hiom Debra - Gateway Standardization: A Quality Assurance Framework For Metadata

Kenney R. and Warden R., An Open Access Future? Report from the eurocancercoms project, ecancermedicalscience, DOI: 10.3332/ecancer.2011.223 Redwood City, San Francisco Bay, USA, September 10 to September 12, 2009

Kitchenham Barbara, Shari Lawrence Pfleeger Software Quality: The elusive target Systems/Software, Inc. IEEE SOFTWARE 0740-7459/96 1996 IEEE Vol. 13, No. 1: JANUARY 1996, pp. 12-21

Lagoze C. and Van de Sompel Herbert. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html

Lagoze Carl, Herbert, Michael Nelson, 2002. Implementing Guidelines for the Open Archives Initiative for Metadata Harvesting: Guidelines for Harvesting Implementes http://www.openarchives.org/OAI/2.0/guidelinesharvester.2002-06-10.htm

Lawrence, S., Giles, C. (1999). Accessibility of Information on the Web. Nature, 400, 107-109

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

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

NISO A Framework of Guidance for Building Good Digital Collections (Bethesda, MD:NISO Press, 2007),61-2

NISO Understand Metadata http://www.niso.org/publications/press/UnderstandingMetadata.pdf NISO Press 2001

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

Ochoa Xavier, Duval Erik - Quality Metrics for Learning Object Metadata

OCLC/RLG Working Group on Preservation Metadata - A Metadata Framework to Support the Preservation of Digital Objects Report - a http://www.oclc.org/research/pmwg/ June 2002

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

Pipriani Baba and Ernst Denise– A Model for Data Quality Assessment

Salo, Dorothea. "Name Authority Control in Institutional Repositories." *Cataloging & Classification Quarterly* 47, no. 3/4 (2009): 249-261. http://digital.library.wisc.edu/1793/31735

Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Communications of the ACM 40(5) (1997) 103–110

Stvilia – Measuring Information Quality Dissertation – Urbana-Illinois, 2006

Stvilia Besiki, Gasser Les, Twidale Michael B., Shreeves Sarah L., Tim W. Cole - Metadata Quality for Federated Collections

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)

Van de Sompel H., Nelson M. L., Lagoze C., Warner S.: Resource Harvesting within the OAI-PMH Framework – D-Lib Magazine December 2004 – Volume 10 N. 12 ISSN 1082-9873 - http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html

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

Wang, R, Strong D – Beyond accuracy: What data quality means to data consumers – journal of Management Information System 12(4), 5-35

Ware, Mark. Pathfinder Research on Web-based Repositories – Final Report. Publisher and Library/Learning Solutions, 2004

Xiaoming Liu et al. « Arc - An OAI Service Provider for Digital Library Federation" D-Lib Magazine April 2001 Volume 7 Number 4 ISSN 1082-9873

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

Zubrow David - Can you trust your data? Measurement and Analysis Infrastructure Diagnosis 2007 SEI

[5] C. Lagoze and H. V. de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002. http://www.openarchives.org/OAI/openarchivesprotocol.html.

[18] Ede, S.: Fitness for purpose: The future evolution of bibliographic records and their delivery. Catalogue & Index 116 (1995)

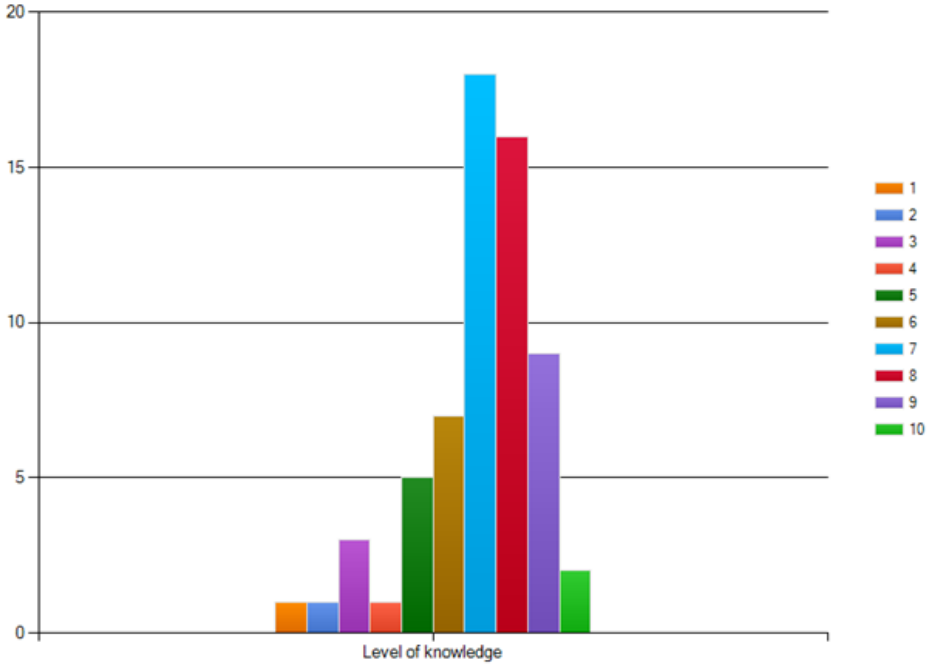
[36] Baden Hughes, Metadata Quality Evaluation: Experience from the Open Language Archives Community. 7th International Conference on Asian Digital Libraries, ICADL 2004, Shanghai, China, December 13-17, 2004. Proceedings, pp. 320-329.

[78] Alain Abran, Rafa Al Qutaish, Jean-Marc Desharnais, Naji Habra - An Information Model for Software Quality Measurement with ISO Standards

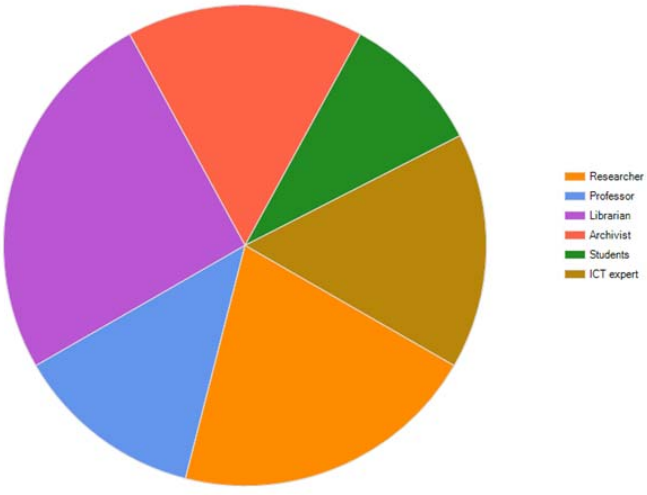
ANNEX I

Survey Results

1. Please, evaluate your level of knowledge of the Dublin Core



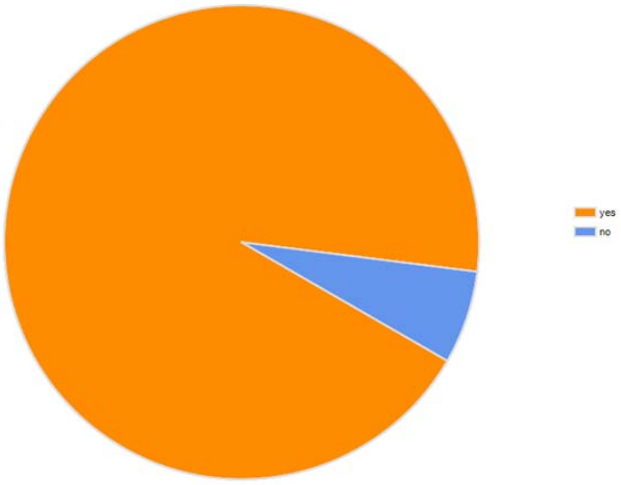
2. Please indicate your profile



Researcher - 20,6%
Professor - 12,7%

Librarian - 25,4%
Archivist - 15,9%
Students - 9,5%
ICT expert - 15,9%

3. Does your work/activity also include the definition and use of metadata?



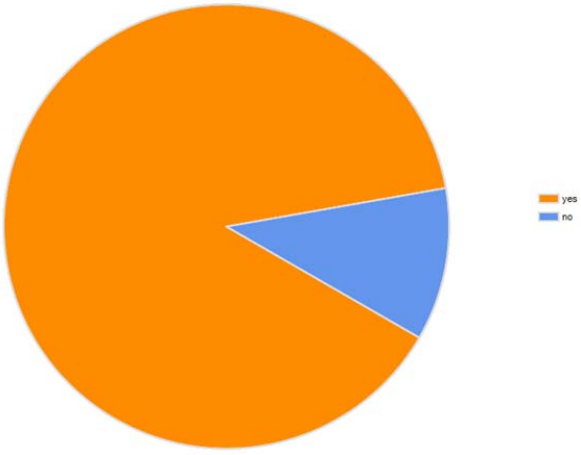
Yes: 93,7%

No: 6,3%

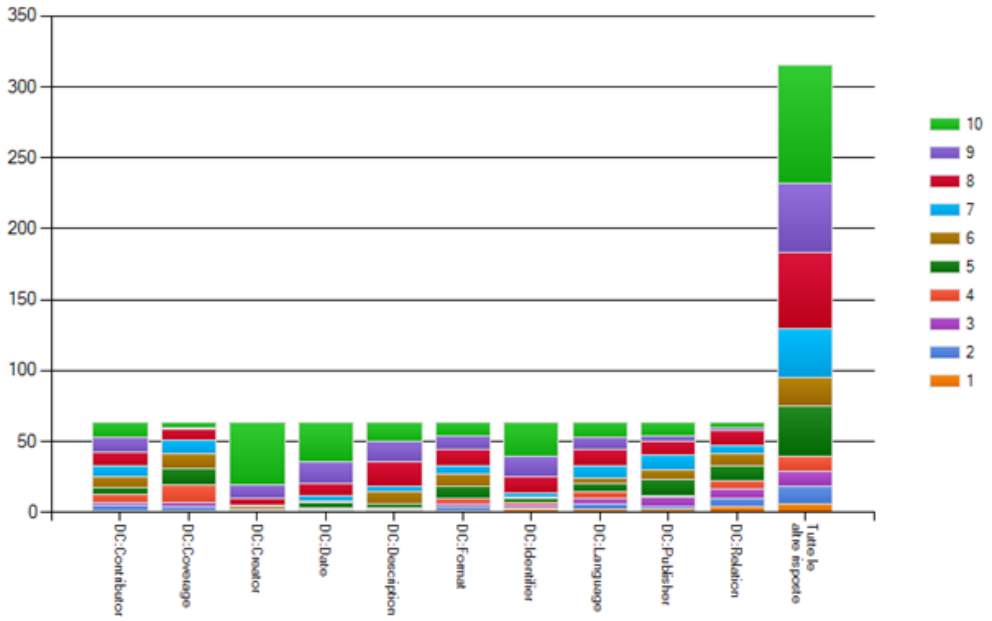
4. Have you ever dealt with the Quality of metadata?

Yes: 88,9%

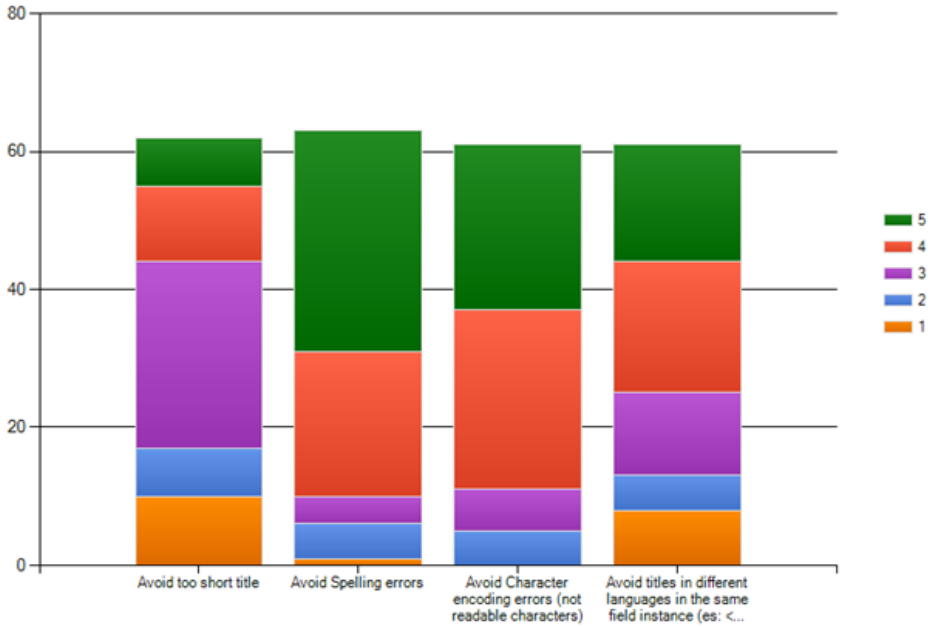
No: 11,1%7



5. We consider a DC field complete if it is NOT EMPTY in the metadata record. In order to estimate the contribution of each DC field in determinating the COMPLETENESS of a metadata record in the OA Institutional Repositories, we ask you to assign a WEIGHT to each 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)



6. We consider accurate a value in a metadata field if it is compliant with the standards defined for that field. Please indicate the level of importance of the following recommendations to evaluate the ACCURACY of the value in the <DC:Title> field 1 (low importance) to 5 (Max importance)



7 Please indicate the level of importance of the following recommendations to evaluate the ACCURACY of a value in the <DC:Description> field 1 (low importance) to 5 (Max importance)