A Method for Gathering and Classification of Scientific Production Metadata in Digital Libraries

Elisabete erreira and Marcos Sfair Sunye
Department of Computer Science, Federal University of Paraná, Curitiba, Brazil

ey ords Digital Libraries Metadata etrie al Digital epositories Scientific ournal Publication pen Access

Abstract his paper introduces a methodology for the automatic loading of metadata and open access scientific articles spread out in scientific ournals in Institutional Digital epositories ID s obtained through information s e traction from the researchers curricula A further ob ecti e is to help the institution for planning the costs re uired to support the gro th of their digital en ironment considering the scientific data that ould be stored in it he aggregation of scientific production in a single institutional digital en ironment allo s institutions to generate internal indicators of scientific and technological production conduct studies through the application of data mining tools as ell as support the implementation of management policies or the purpose of implementation a set of components as de eloped for collecting scientific articles free of all restrictions on access

1 INTRODUCTION

igh education institutions are responsible for the ma ority of the scientific production in the form of published ournal articles reports conference papers and so forth

Although academic institutions are the ma or scientific no ledge producers the tas s of aggregating and uantifying the no ledge produced by their researchers is a difficult tas Setenares i et al 2016 herefore setting specific criteria for planning and distributing resources to encourage scientific production by their faculties become a rele ant goal

Another issue for the institutions is monitoring their intellectual producti ity through indicators and the proper planning of the archi ing and preser ation process of digital materials o er the long term his situation ta es place due to the lac of a tool that effecti ely determines the costs of implementation and maintenance of their digital en ironments

Digital Libraries that are incontestably rele ant no adays are responsible for aggregating selecting structuring offering access interpreting distributing and preser ing items of intellectual resources of an institution hence these components are financially accessible to the community Langiano 2005 In addition by increasing access to research results of an institution Digital Libraries benefit professionals and students ho use their resources in teaching and learning tas s

he results of the no ledge produced in the form of scientific articles by an institution are published in scientific ournals hich are considered the fastest and most affordable ay to disseminate scientific information the findings of research or and hat these or s represent to the community rofman 2012

e ertheless digital libraries fail in har esting selection and aggregation of items of scientific production published in periodical ournals or instance many of them focus solely on pro iding the scientific production of their educational programs in the form of monographs theses and dissertations

here are many methods to populate ID s such as self-submission and semi-automated mechanisms Another form is har esting by AI-PM a ailable at http openarchi es org pmh a protocol created to promote interoperability bet een libraries and digital repositories as an effort to impro e access to e-print archi es in order to increase the a ailability of scholar communication o e er all of them re uire the in ol ement of the authors and or the library staff

AI-PM defines hich criteria must be met to facilitate the efficient dissemination of content in digital en ironments In its conte t there are t o types of pro iders hich re uire both the publishers and digital repositories to support it and ha e it enabled

- data pro iders hich are repositories that e pose their structured metadata according to AI-PM and
- ser ice pro iders hich ma e ser ice re uests

Ferreira, E. and Sunye, M.
A Method for Gathering and Classification of Scientific Production Metadata in Digital Libraries.
DOI: 10.5220/0006358203570364
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 357-364
ISBN: 978-989-758-247-9

ia AI-PM to har est the a ailable metadata

egarding this scenario the fact that scientific production of institutional articles end up scattered in scientific ournals consists a serious problem In this sense access and identification of this scientific no ledge by the community and e en by the institution itself that produced it is often hampered Li e- ise institutions also lac information about ho much of the teaching staff is a are of the a ailability of their science production free of all restrictions on access Moreo er another difficulty is identifying open access articles hen this information is not found on metadata

In the field of scientific publication *open access* means publications on the Internet that allo reading copying distribution or re-use for la ful purposes ithout technical financial or legal barriers as ell as guaranteeing the author s moral and patrimonial rights pen Society Institute 2002 he philosophy behind open access is a trend that has been obser ed in recent years to ards the use of tools strategies and methodologies to communicate ne scientific research

In this conte t this article proposes a methodology for loading open access articles on digital repositories obtained through information e traction from curricula of an institution s researchers

ra ilian researchers ha e their scientific production registered at an academic national database the Lattes Platform a ailable at http lattes cnp br

or implementation purposes a study di ided into 4 parts as reali ed 1 gathering and processing of metadata from a researchers curricula database 2 de elopment of a script for collecting open access scientific articles 3 selection of a soft are for loading and con erting metadata to the Dublin Core format and 4 populating a digital repository by importing the ac uired data or this study purposes the ID of a ra ilian institution as used

his or is organi ed as follo s the second section tal s about metadata its characteristics and importance in inde ing digital ob ects in digital en ironments persistent identifiers li e D Is and handles their uses and finalities for preser ation of digital ob- ects on the long term are conte tuali ed in the third section the fourth section brings the concepts about Digital Libraries such as their history importance and characteristics the fifth section describes the proposed method and analy es its application and finally the main points of this paper are summari ed and suggestions for future or s are presented in the conclusion

2 METADATA

Metadata are information related to a stored resource either physical or not that not only identify and describe it but also document its beha ior function and use as ell as its relationship to other digital ob ects and ho it should be managed Metadata are structured in the form of te t and ey ords and generally contain direct information such as author name creation date sub ect but can also be comple and harder to define as the opinion consensus of arious people on the same boo Langiano 2005 hus metadata pro e to be essential to facilitate disco ery of rele ant content in digital libraries

urthermore an item or ob ect a ailable in digital media should sur i e the successi e generations of hard are and soft are Gi en such comple ity and the importance in designing digital ob ects metadata a study as proposed to categori e them into fi e types aca 1998

- Administrati e used in the management and administration of information resources such as ersion control and copyright information
- echnical related to the operation or beha ior of system metadata for e ample scanning processes
- Descripti e used to describe and identify resource information for e ample speciali ed inde es and search aids
- Preser ation related to the preser ation of information resources for e ample policies relating to the bac up of digital ob ects
- se related to the le el and type of use of information resources

In this article descripti e metadata are used for identification of bibliographic content of scientific or s

2.1 Metadata Schema

Metadata schemas are sets of elements designed for a specific purpose that are used to describe an information resource he elements definitions or meanings are no n as the schema s semantics and the alues of a gi en element are its contents Metadata schemas generally specify the names of elements and the corresponding semantics Sayao 2007b Metadata should be carefully planned and support interoperability ith other digital libraries hence facilitating the location and use of digital ob ects Metadata schemas and metadata standards e ist to enable the effecti e sharing of resources bet een institutions and users

2.1.1 Dublin Core - DC

Dublin Core is a metadata schema proposed in 1995 to promote metadata interoperability DCMI 2012 Dublin Core uses a set of simple but effecti e elements that describe a ide ariety of net or resources and hose semantics ere established by an international consensus of professionals from arious disciplines such as library science computing te t mar up community museums and other related fields LA A 2002 his metadata schema uses fifteen descripti e elements standardi ed by technical ocabularies and specifications maintained by the Dublin Core Metadata Initiati e DCMI 2012 Dublin Core is the metadata schema adopted by the analy ed ID

he Dublin Core elements are identified as dc and ha e a single alue Since each element has unlimited occurrence ualifiers are used in order to distinguish the alue of each occasion hich may ha e an identifier called schema or modifier Al es and Sou a 2007 according to the synta **dc.*element.qualifier*** as sho n in igure 1 Although this schema pro ides an element for identifying rights it is not commonly included in article s metadata

Element	Value	Language
dc.contributor.author	Barbosa, Eduardo Mayer	
dc.contributor.author	Rodrigues, Tamires Maria	
dc.date.accessioned	2015 06 13T00:31:47Z	
dc.date.available	2015 06 13T00:31:47Z	
dc.date.issued	2015 06 12	
dc.identifier.uri	http://hdl.handle.net/1884/38213	
dc.language.iso	pt_BR	pt_BR
dc.rights	Attribution 3.0 United States	*
dc.rights.uri	http://creativecommons.org/licenses/by/3.0/us/	*
dc.subject	Esporte de Orientação, Leitura de Mapas, Geografia, Ensino.	pt_BR
dc.title	USO DO ESPORTE DE ORIENTAÇÃO EM AMBIENTE REDUZIDO PARA O ENSINO DE LEITURA DE MAPAS	pt_BR
dc.type	Working Paper	pt_BR

igure 1 E amples of Dublin Core elements

2.2 Metadata and Interoperability

he information of digital ob ects stored in Digital Libraries or epositories is called content and is di ided into data and metadata hile the first corresponds to the generic term that describes the information in digital format the second is data about the data itself Langiano 2005 Metadata if carefully constructed bring se eral ad antages for users of digital libraries since a standardi ed representation of the a ailable information resources in electronic form pro ides a broad and accurate access to the content stored in these en ironments

Pro ided that digital ob ects must sur i e successi e generations of hard are soft are and systems metadata pro e to be ital by allo ing them to e ist independently of the system in use for storage and search In this sense metadata are essentially technical descripti e and should be preser ed in order to document the creation and maintenance of a digital ob ect as ell as its a ailability and relationships ith other ob ects or digital ob ects to remain accessible and intelligible o er time the transportation and preser ation of their metadata must be possible aca 1998 o facilitate the search and access to the digital ob ects contents a metadata schema is selected to describe the arious e isting types of contents e g ideos sounds images te ts ebsites etc according to the library or repository s purpose

2.3 Metadata Harvesting

Se eral studies on the creation and updating of digital en ironments as ell as best practices to be follo ed by institutions are found in literature amos et al 2012 Institutional digital en ironments promote and contribute to the dissemination of scientific production once they are one of the tools that guarantee the isibility of the institution and its researchers herefore they should al ays be a ailable and constantly updated hus for the success of a digital library the effecti e interaction bet een the de elopment and maintenance team and the staff responsible for its archi es is a rele ant issue Leite 2009 he digital repositories analy ed for this or are populated by the forms belo

2.3.1 Automation by OAI-PMH

he use of the AI-PM for automatic update of metadata in digital libraries is based on the metadata e traction from national and international databases and faces the lac of standards for e traction of open access scientific publications metadata urthermore ust the scientific articles found in these databases hether open access or not that meet the database s specific guidelines are sufficient to ma e the locating of all articles produced by an institution be ineffecti e

2.3.2 Automation by Self-submission

he libraries and ID s identified in this study use the self-submission process to collect their scientific production Self-submission offers in some specific cases support from a library team or customi ed tools for metadata retrie al beyond the indispensable par-

ticipation of authors for gathering information about the licensing of the scientific production in ol ed

ut of the 1 000 top institutional repositories ran ed by ebometrics a ailable at http repositories ebometrics info in the second half of 2014 887 ha e scientific articles 103 ha e only monographs dissertations and theses and 10 ere una ailable rom the 887 that ha e scientific articles 764 use the process of direct self-submission 17 use self-submission ith supporting tools for metadata retrie al 32 ha e submissions made by the library after recei ing data and metadata sent by the authors 27 ha e self-submission made by the authors ith the support of a library 16 ha e automated collection by AI-PM and in one of them a library staff identifies the scientific production from the authors email addresses and later re uests them to submit the publication in the institution s digital library

he aggregation of scientific literature in a single institutional digital en ironment allo s the institution to de elop internal indicators of scientific and technological production carry out research by applying data mining tools and support the implementation of management policies

3 PERSISTENT IDENTIFIERS

he archi ing and preser ation of digital materials o er the long term is a difficult and e pensi e tas that re uires substantial resources and institutional commitment IS 2007 In the mid 1990s ith the orld ide eb s populari ation there ere persistent identifiers that corresponded to uni ue identification elements added to digital ob ects hich regardless of their location or format ensured that they ere accessible in the long term despite physical and technological changes Sayao 2007a Persistent identifiers are typically found as s niform esource ame Cs niform esource Characteristics and Ls niform esource Locators among others Sollins and Masinter 1994

3.1 The Handle System

he *Handle System*○ persistent identifier as de eloped in 1994 by the Corporation for ational e-search Initiati es C I in the nited States It is a component of the digital ob ects architecture hich pro ides a safe efficient and e tensible resolution to uni ue and persistent identifiers esolution ser ices are the mechanisms by hich a particular persistent identifier is lin ed to an L here the digital ob ect is stored

3.2 Digital Object Identifier - DOI

D I Digital b ect Identifier as presented for the first time at the ran furt oo air in 1997 and a little further in the same year the International D I oundation ID as created to manage the system he D I is a proprietary implementation of the *Handle System* originated from a oint initiati e of three trade associations in the boo industry International Publishers Association International Association of Scientific echnical and Medical Publishers and Association of American Publishers It emerged as a generic frame or for the content ID management through digital net or s International D I oundation 2015 Since then D Is ha e been used to assign and disseminate information on intellectual property rights to the digital ob ects Sayao 2007a

or the correct location of a digital ob ect using D I a minimum of structured metadata such as bibliographic and commercial information should e ist Metadata assigned to a digital ob ect gi e the user the assurance that the resource found is effecti ely hat he or she is loo ing for he data model used by a D I identifier pro ides a conte tual metadata system that supports interoperability bet een different e isting metadata schemas in a digital en ironment his model consists of an interoperable data dictionary plus an underlying structure for applications

4 DIGITAL LIBRARIES

he materials or digital ob ects a ailable in a digital library may deri e from digital copies of e isting materials in physical media for instance boo s prints manuscripts etc and or from ob ects e isting only in digital media such as digital photos e-boo s ideos and others

Aiming at a reliable digital preser ation process it is important that in addition to using rigorous scientific methodology for the generation of no ledge the results obtained by the academic and scientific research of an institution are disseminated in open access digital repositories lin ed to a persistent identifier As a matter of fact persistent identifier is a uni ue name for a digital ob ect that is independent of its location or format ensuring that the ob ect is accessible independent of physical and technological changes Sayao 2007a

he adoption of ID s by uni ersities and research centers promotes an increase in the isibility and

competiti eness of these institutions hich in its ay contributes to scientific de elopment Leite 2009

Institutional repositories can belong to uni ersities laboratories and research institutes hereas the thematic repositories are arranged by no ledge area ithout institutional boundaries he adoption of digital repositories hen ell planned and properly implemented promotes increased isibility of research results the researcher and of the institution itself Leite 2009 ID s a fundamental element of today s digital libraries ha e been hea ily used to promote scientific production from research and teaching acti ities Leite 2009

5 ARCHITECTURE

he ob ecti e of this study as to de elop a set of components for metadata har esting and classification of open access scientific articles in an ID he ueries used D I identifiers to retrie e information from the publication since this is the uni ue permanent identifier to reco er an article in the online en ironment D Is ha e been obtained from the researchers curricula from an institution

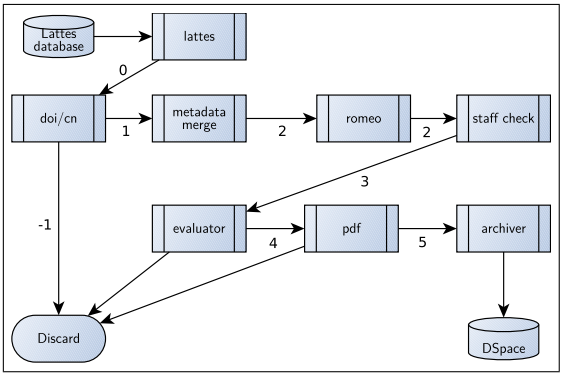
5.1 The Lattes Platform

he Lattes Platform is the e perience of C P in integrating databases of resumes research groups and institutions into a single information system Plataforma Lattes 1999

Currently teachers and researchers from ra ilian institutions ho produce scientific or and participate in go ernmental programs such as the Coordination for the Impro ement of igher Education Personnel Capes a ailable at http capes go br and the ational Council for Scientific and echnological De elopment C P a ailable at http cnp br pagina-inicial are ad ised to inform their scientific productions on this database urthermore it is possible for an educational institution to access the scientific production of its faculties through the Lattes E tractor system he data e traction is pro ided ia ML files containing all the institution s scientific production registered on the platform by research groups teachers researchers and students

5.2 Proposal

Each article is processed by the components in the system and assigned to a specific stage from -1 to 5 as sho n in igure 2



igure 2 Stage o for article selection

he D Is and metadata e tracted from ARTIGO-PUBLICADO *published article* tags in curricula of Lattes Platform are stored in a database ith stage 0

```
1 <ARTIGO-PUBLICADO SEQUENCIA-PRODUCAO="26" ORDEM-IMPORTANCIA="">
2     <DADOS-BASICOS-DO-ARTIGO NATUREZA="COMPLETO" TITULO-DO-ARTIGO="Production of mortadella:
  Behavior of Listeria monocytogenes under commercial manufacturing and storage conditions"
  ANO-DO-ARTIGO="2001" PAIS-DE-PUBLICACAO="" IDIOMA="Inglês" MEIO-DE-DIVULGACAO="IMPRESSO"
  HOME-PAGE-DO-TRABALHO="[http://www.sciencedirect.com/science?_ob=articleurl&amp;_udi=B6T9g-
  4183sdx-2&amp;_user=972052&amp;_coverdate=01%2F31%2F2001&amp;_rdoc=2&amp;_fmt=summary&
  amp;_orig=browse&amp;_srch=doc-
  info(%23toc%235114%232001%23999429998%23210730%23fla%23display%23volume)&amp;_cdi=5114&
  amp;_sort=d&amp;_doi[doi:10.1016/s0309-1740(00)00068-1]" FLAG-RELEVANCIA="SIM"
  DOI="10.1016/S0309-1740(00)00068-1" TITULO-DO-ARTIGO-INGLES="" FLAG-DIVULGACAO-
  CIENTIFICA="NAO"/>
3     <DETALHAMENTO-DO-ARTIGO TITULO-DO-PERIODICO-OU-REVISTA="Meat Science" ISSN="03091740"
  VOLUME="57" FASCICULO="1" SERIE="" PAGINA-INICIAL="13" PAGINA-FINAL="17" LOCAL-
  DE-PUBLICACAO=""/>
4     <AUTORES NOME-COMPLETO-DO-AUTOR="Mariza Landgraf" NOME-PARA-CITACAO="LANDGRAF, Mariza"
  ORDEM-DE-AUTORIA="2" NRO-ID-CNPQ="3040125005832575"/>
5     <AUTORES NOME-COMPLETO-DO-AUTOR="Bernadete Dora Gombossy de Mello Franco" NOME-PARA-
  CITACAO="FRANCO, Bernadete Dora Gombossy de Mello" ORDEM-DE-AUTORIA="3" NRO-ID-
  CNPQ="4629835197940139"/>
6     <AUTORES NOME-COMPLETO-DO-AUTOR="Maria Teresa Destro" NOME-PARA-CITACAO="DESTRO, Maria
  Teresa" ORDEM-DE-AUTORIA="4" NRO-ID-CNPQ="4460416634163824"/>
7     <AUTORES NOME-COMPLETO-DO-AUTOR="Luciano dos Santos Bersot" NOME-PARA-CITACAO="BERSOT, L.
  S.;BERSOT, L.D.S;Bersot, Luciano dos Santos;BERSOT, L." ORDEM-DE-AUTORIA="1"/>
8 </ARTIGO-PUBLICADO>
```

igure 3 E ample of an ARTIGO-PUBLICADO tag in the Lattes ML

henceforth the metadata of each D I is re uested from the *dx.doi.org* resol er as sho n in igure 4 In this step the D Is are also alidated being stored ith stage 1 in alid D Is are assigned for deletion ith stage -1

```
  "DOI": "10.1007/s00799-012-0080-5",
  "type": "journal-article",
  // [...]
  "ISSN": ["1432-5012", "1432-1300"],
  "title": "Extending OAI-PMH over structured P2P networks for digital preservation",
  "container-title": "Int J Digit Libr",
  "publisher": "Springer Science + Business Media",
  "volume": "12",
  "issue": "1",
  "issued": {
    "date-parts": [[2012, 2, 28]]
  },
  "author": [
    { "given": "Everton F. R.", "family": "Seára", "affiliation": [] },
    { "given": "Luis C. E.", "family": "Bona", "affiliation": [] },
    { "given": "Tiago", "family": "Vignatti", "affiliation": [] },
    { "given": "Andre L.", "family": "Vignatti", "affiliation": [] },
    { "given": "Anne", "family": "Doucet", "affiliation": [] }
  ]
}
```

igure 4 D I metadata in the *citeproc* format

hereafter the sets of metadata retrie ed from the Lattes and D I bases are merged ith D I metadata ta ing precedence and being mar ed as stage 2

According to the site S E PA oME a ailable at http sherpa ac u romeo an initiati e for

identification of scientific publications according to the open access mo ement the publishing ournal s International Standard Serial umber ISS is eri- fied according to the open access mo ement as sho n in igure 5

```
<!-- [...] -->
<romeoapi version="2.9.9">
  <!-- [...] -->
  <journals>
    <journal>
      <jtitle>Physical Review A</jtitle>
      <issn>1050-2947</issn>
      <zetocpub>American Physical Society</zetocpub>
      <romeopub>American Physical Society</romeopub>
    </journal>
  </journals>
  <publishers>
    <publisher id="10">
      <name>American Physical Society</name>
      <!-- [...] -->
      <preprints>
        <prearchiving>can</prearchiving>
        <prerestrictions />
      </preprints>
      <postprints>
        <postarchiving>can</postarchiving>
        <postrestrictions />
      </postprints>
      <pdfversion>
        <pdfarchiving>can</pdfarchiving>
        <pdfrestrictions />
      </pdfversion>
      <!-- [...] -->
      <romeocolour>green</romeocolour>
    </publisher>
  </publishers>
</romeoapi>
```

igure 5 esult of a uery to S E PA oME

In the case of production restricted to the institution at stage 2 there as a uery to institutional staff database Similar names are resol ed by an algorithm for homonyms that compares initials in the author s name

- MA CI SA S SIL A M SA S SIL A MA CI S SIL A MA CI SIL A M S SIL A e M SIL A
- MA ICI SIL A e M SIL A

Gi en the e ample abo e articles authored as M SIL A ould be discarded since the author s name is ambiguous In this step articles ith ISS produced by an institutional author throughout his or her permanence at the institution and that can be published according to the open access mo ement are set up as stage 3

ereafter all articles that can be published in digital repositories are set as stage 4

In order to obtain the articles full te t as PD algorithms to search the periodical s ML pages ere de eloped hen the articles PD s are stored in a database In this scenario some difficulties emerged hile some periodicals allo the access to the PD only by the D Is others ha e loc s by robots against har esting or other reasons or each difficulty a specific algorithm as de eloped

he articles PD s a ailable for storage according to the open access mo ement ere stored in database ith stage 5

nce the metadata and PD s ha e been selected a directory structure in the Simple Archi e ormat as de eloped for importing by DSpace soft are as sho n in igure 6

```
<dublin_core>
  <dcvalue element="identifier" qualifier="other">10.1073/PNAS.0508170103</dcvalue>
  <dcvalue element="title">Linoleic acid hydroperoxide reacts with hypochlorous acid, generating peroxyl radical intermediates and singlet molecular oxygen</dcvalue>
  <dcvalue element="relation" qualifier="ispartof">Proceedings of the National Academy of Sciences, v. 103, n. 2</dcvalue>
  <dcvalue element="publisher">Proceedings of the National Academy of Sciences</dcvalue>
  <dcvalue element="date" qualifier="issued">2005-12-30</dcvalue>
  <dcvalue element="identifier" qualifier="issn">0027-8424</dcvalue>
  <dcvalue element="identifier" qualifier="issn">1091-6490</dcvalue>
  <dcvalue element="contributor" qualifier="author">D. Rettori</dcvalue>
  <dcvalue element="contributor" qualifier="author">G. R. Martinez</dcvalue>
  <dcvalue element="contributor" qualifier="author">M. H. G. Medeiros</dcvalue>
  <dcvalue element="contributor" qualifier="author">O. Augusto</dcvalue>
  <dcvalue element="contributor" qualifier="author">P. Di Mascio</dcvalue>
  <dcvalue element="contributor" qualifier="author">S. Miyamoto</dcvalue>
</dublin_core>
```

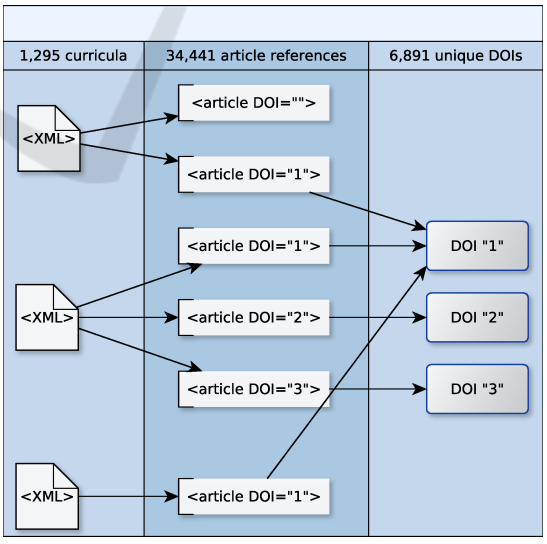
igure 6 E ample of Metadata Archi e according to the Simple Archi e ormat

inally the selected articles PD s and their respecti e metadata ere stored at the ID

5.3 Case Study and Analysis

he analy ed repository part of the Digital Libraries of the ederal ni ersity of Parana in ra il as established in 2004 using the DSpace platform his digital en ironment presents collections of different types of scientific output such as theses monographs and dissertations among others o e er the repository currently does not include scientific articles

Since an article can be referenced by more than one curriculum these references ere unified by their D Is igure 7 hen a ailable



igure 7 Count of normali ed metadata

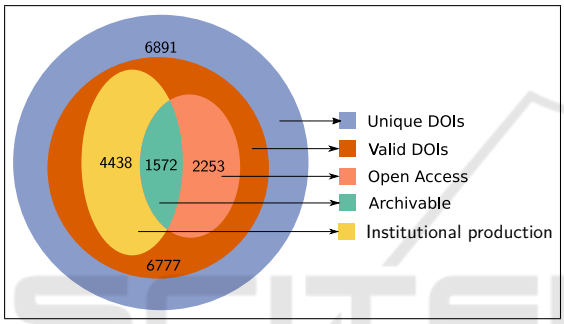
In the tag Published Article from Lattes Platform s ML 1 295 occurrences ere obtained hich resulted in 34 441 references to scientific articles rom these references 8 969 ere classified ith

D I attribute 6 891 ith uni ue D Is and after D I s resol er submission 6 777 ith alid D I identifier

he alid D Is referenced 36 463 authors of hom 8 907 ere researchers from the analy ed institution

f the 2 783 ISS s found 545 19 6 belong to open access periodicals 2 029 42 9 belong to non open access periodicals and 209 7 5 are not listed at S E PA oME

f the 6 777 articles ith alid D I 2 253 allo archi ing in digital repositories and 4 438 ere produced by the institution s researchers 1 572 articles meet both criteria and can be archi ed at the Institutional Digital epository his scenario is illustrated by igure 8



igure 8 Count of open access articles identified

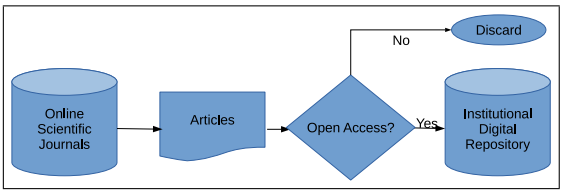
In order to ensure the identification of only the scientific production of the uni ersity s faculties but also considering the possible production of a professor currently inacti e but engaged in another institution cross-references ere made bet een the information a ailable on the Lattes Platform and the institution s administrati e database hich resulted in 1 241 classified professors

After ards the metadata of articles by selected professionals ere retrie ed from Cross ef a ailable at http crossref org a registry authority for D Is by a script hich found 2 293 ournals he metadata ere stored in a relational database

It as necessary to chec the license under hich the articles ere produced his as accomplished by uerying the publishing ournal s ISS on S E PA oME

he article metadata obtained and stored in the database ere cross-referenced ith the list of open access ournals resulting in 2 287 articles published under open access and hose published ersion could be redistributed by the institution s digital libraries

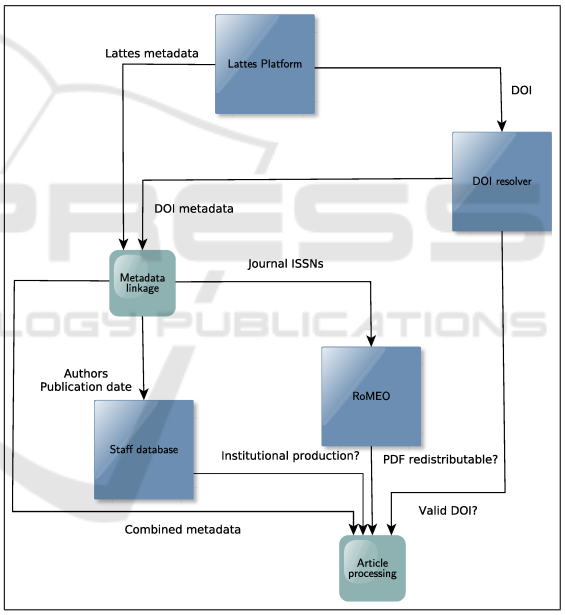
igure 9 sho s the selecti e process of the scientific production of articles classified for importation in the digital repository



igure 9 Selection of scientific articles for importation

nce a scientific article a ailable under open access is identified its metadata plus a specific metadata item hich indicates the license under hich the article is a ailable are collected and stored in a database allo ing the identification of these items in a digital repository he purpose of such action as to add the data for subse uent metadata transference according to the *Dublin Core* metadata format as it is the schema adopted by the ID that is ob ect of this study

he diagram in igure 10 sho s the o and the entities and decisions in ol ed



igure 10 lo entities and relationships in the selection of the institution s open access scientific articles

he process of automated loading follo ed these steps 1 creation of a community called Scientific Production 2 creation of a sub-community called Articles and 3 population of the data and metadata

5.3.1 Result Analysis

Scientific publications stem from research pro ects hey aim to disseminate scientific research to the community in order to allo others to use it and e aluate it in other ie s rofman 2012

Although the authors of scientific or s are as ed to pro ide their or in a standardi ed electronic format on at least one open access repository by the erlin Declaration MPG 2003 it is still clear the lac of a areness by many of the authors related to this topic his research identified that 35 4 of the institution s scientific production ere produced in open access format

6 CONCLUSION

he adoption of ID s promotes the dissemination of technical and scientific content produced by an institution and culturally enrich those ho benefit from it Aggregation of scientific production in one location enables access to a great amount of information and therefore encourages the transfer of no ledge his or also demonstrated ho important it is for the uni ersity staff to o n and maintain their data updated in a curricula database as ell as properly inform the D Is associated ith their publications since this is the only permanent identifier of an article for reco ery in the eb Moreo er it as sho n that the institution could plan the costs re uired to maintain its digital en ironment by determining the olume of scientific production to be stored in its digital library In this scenario the institution ould also be able to measure the real impact produced by their academic community

his study aimed at de eloping a set of components for collection and classification of metadata of scientific articles produced under open access in an ID ithout assigning to their researchers the tas of eeping their curricula data up to date in curricula database hus other inds of scientific production ould be orth being classified and aggregated in the Digital Library in future de elopments such as the classification and selection of metadata of 1 scientific e ents 2 uni ersity e tension acti ities

REFERENCES

Al es M and Sou a M 2007 Estudo de correspondencia de elementos metadados Dublin core e marc 21 *RDBCI* 4 2

aca M 1998 *Introducción a los metadatos: vías a la información digital* Getty Information Institute

rofman P 2012 A importancia das publicacoes cient ficas *Cogitare Enfermagem* 17 3 419

DCMI 2012 Dublin Core Metadata Element Set ersion 1 1 A ailable at http dublincore org documents 2012 06 14 dces Accessed in 2015-11-11

International D I oundation 2015 D I handboo A ailable at http doi org doi_handboo 1_Introduction html 1 2 Accessed in 2015-06-18

Langiano d C 2005 m Mecanismo para Automatiar a Criacao dos Metadados das Imagens de ibliotecas Digitais e Pro er uscas por Conteudo Master s thesis P

LA A 2002 u est-ce ue le dublin core A ailable at http lara inist fr lara Accessed in 2015-05-28

Leite C 2009 *Como gerenciar e ampliar a visibilidade da informação científica brasileira: repositórios institucionais de acesso aberto.* I IC

MPG 2003 erlin Declaration on pen Access to no ledge in the Sciences and umanities

IS 2007 A frame or of guidance for building good digital collections A ailable at http niso org publications rp frame or 3 pdf Accessed in 2015-06-17

pen Society Institute 2002 udapest pen Access Initiati e A ailable at http budapestopenaccessinitiati e org Accessed in 2016-10-06

Plataforma Lattes 1999 A ailable at http memoria cnp br eb portal-lattes sobre-a-plataforma Acessado em 2015-03-07

amos C Andretta P I S and Sil a E G 2012 Consideracoes acerca do processo de alimentacao de repositorios atra es da importacao de registros de bases de dados internacionais *RDBCI* 10 1

Sayao L 2007a Interoperabilidade das bibliotecas digitais o papel dos sistemas de identificadores persistentes P L D I andle System Cross ef e pen L *TransInformação* 19 1 65 82 doi 10 1590 s0103-37862007000100006

Sayao L 2007b Metadados para preser acao digital aplicacao do modelo AIS A ailable at http documentoseletronicos ar ui onacional go br Media publicacoes ctdemetadadospreser acao digitalsayao pdf

Setenares i L E Shima Sunye M S and Peres L M 2016 pen digital repositories he mo ement of open access in opposition to the oligopoly of scientific publishers In *Proceedings of the 18th International Conference on Enterprise Information Systems - Volume 2: ICEIS* pages 583 593

Sollins and Masinter L 1994 C 1737 unctional re uirements for uniform resource names A ailable at https tools ietf org html rfc1737 Accessed in 2017-03-03