26th INTERNATIONAL SYMPOSIUM

ETD 2023 NDLTD

Theme: Enriching ETDs and their reach

EXPLORING METADATA QUALITY FOR SCIENTIFIC DATA IN INDIAN RSEARCH DATA REPOSITORIES: A SURVEY

Dr. Sanghamitra Dalbehera

Librarian, Siskha 'O' Anusandhan University,
Bhubaneswar, Odisha

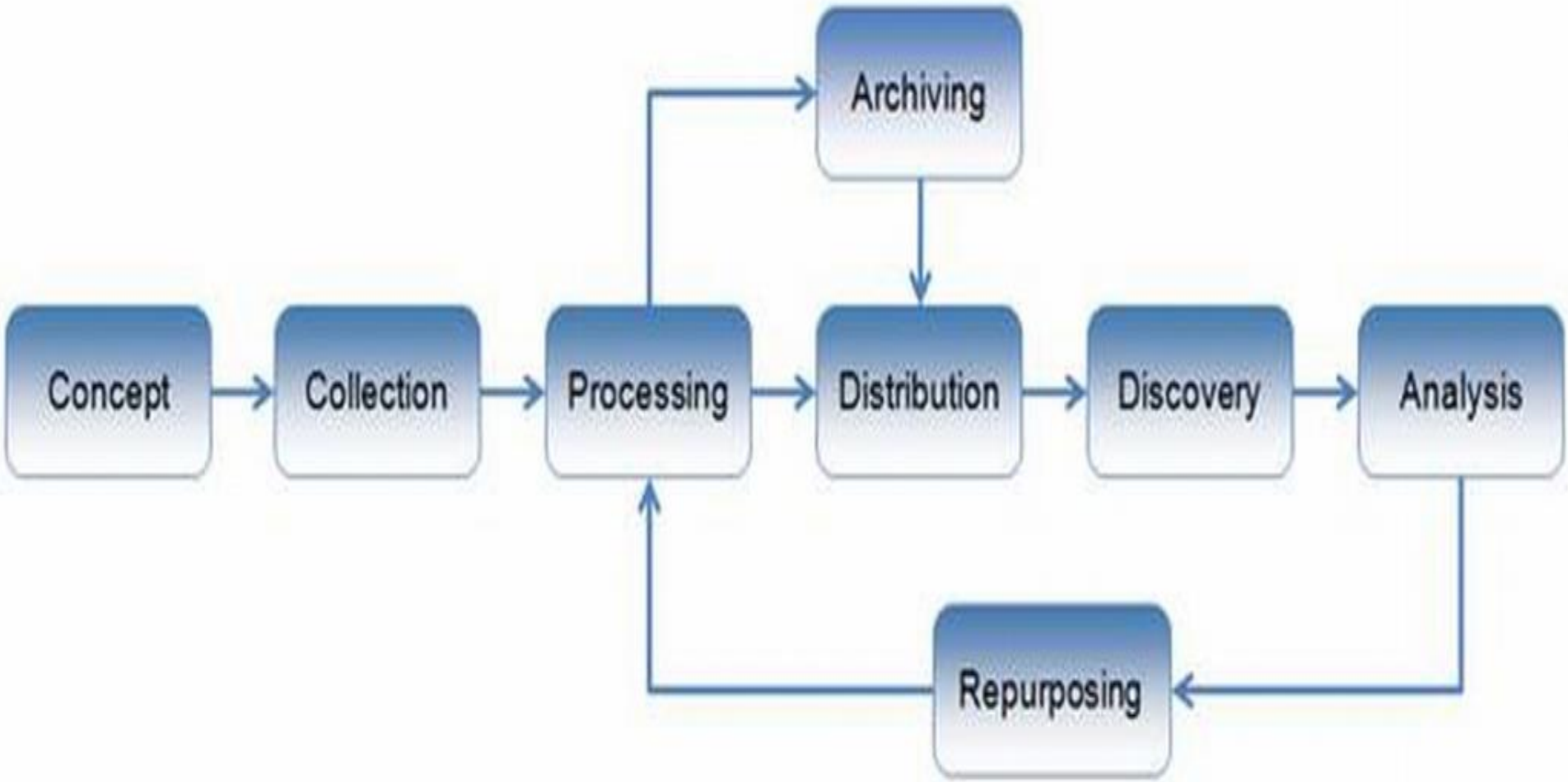
CONTENT

- Background
- re3data.org
- Research Data Repositories
- RDR Registration Workflow of re3data.org
- Typology of RDR
- Analysis of Metadata Types
- DataCite Metadata Elements
- Conceptual framework of Metadata Quality
- Objectives of the study
- Methodology
- Results of the analysis
- Conclusion

BACKGROUND

- Scientific research data is defined as digital data being a (descriptive) part or the result of a research process. This process covers all stages of research, ranging from research data generation, which may be in an experiment in the sciences, an empirical study in the social sciences or observations of cultural phenomena, to the publication of research results.
- Publication of research data, as an independent information object, through a repository.
- Publication of research data with textual documentation as a so-called data paper .
- Publication of research data as enrichment of an interpretive text publication (“enriched publication”)

LIFECYCLE PROCESS IN THE SCIENTIFIC RESEARCH DATA



re3data.org

- is a global registry of research data repositories
- covers research data repositories from all academic disciplines
- helps researchers, funding bodies, publishers and scholarly institutions to find research data repositories
- aims to promote a culture of sharing, increased access and better visibility of research data
- It provides additional information on its service
- It provides information on the terms of access to its data, database and upload
- It provides terms of use and licenses of the data
- It uses a persistent identifier system to make its provided data persistent, unique and citable.

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

Home | **Search** | Browse | Suggest | FAQ | About | Schema | API | Contact | Imprint

Search for Repositories (1234 Reviewed Repositories)

Search

Subject

Add subjects

Content Type

Add content types

Country (of the responsible institutions)

Add countries

Certificates | Open Access | Persistent Identifier

Reset filter

1234 results (1 – 25) Sort by weight

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27

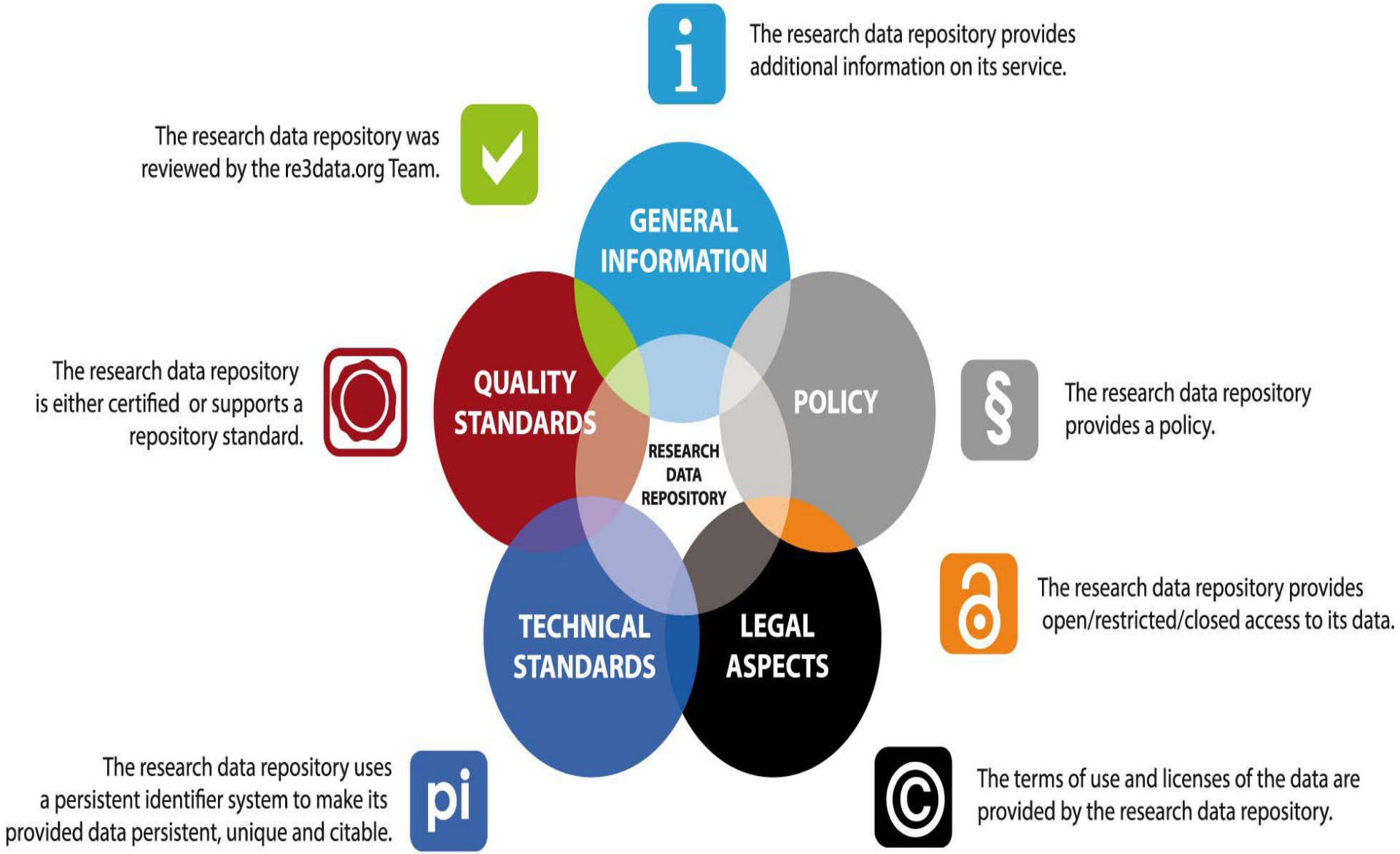
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 »

3TU.Datacentrum

3TU.DC

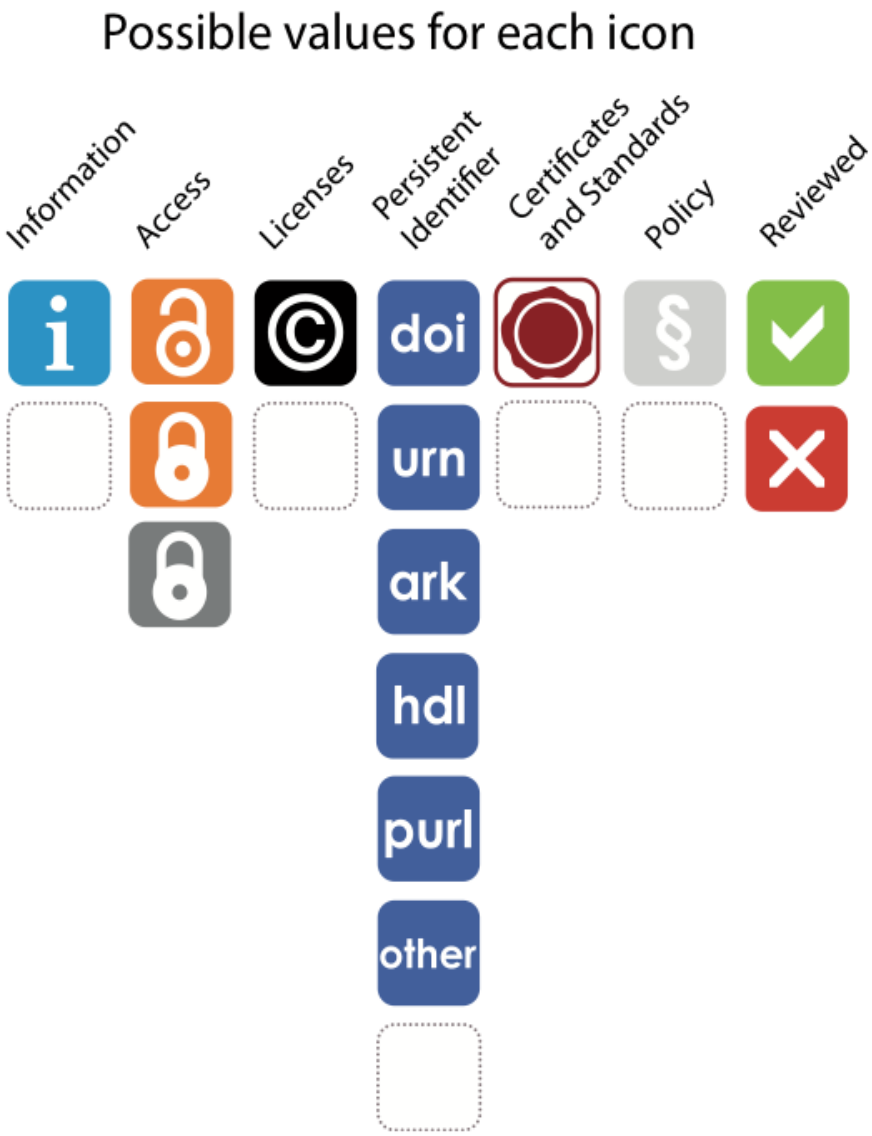
Subjects: Agriculture, Forestry, Horticulture and Veterinary Medicine | Agriculture, Forestry, Horticulture and Veterinary Medicine

Research Data Repository with the corresponding icons



QUALITY REQUIREMENTS

- be run by a **legal entity, such as a sustainable institution(e.g. library, university);**
- **Clarify access conditions to the data and repository as well as the terms of use;**
- Have an **English GUI;**
- Have **focus on research data.**



Research Data Repository Registration Workflow of re3data.org

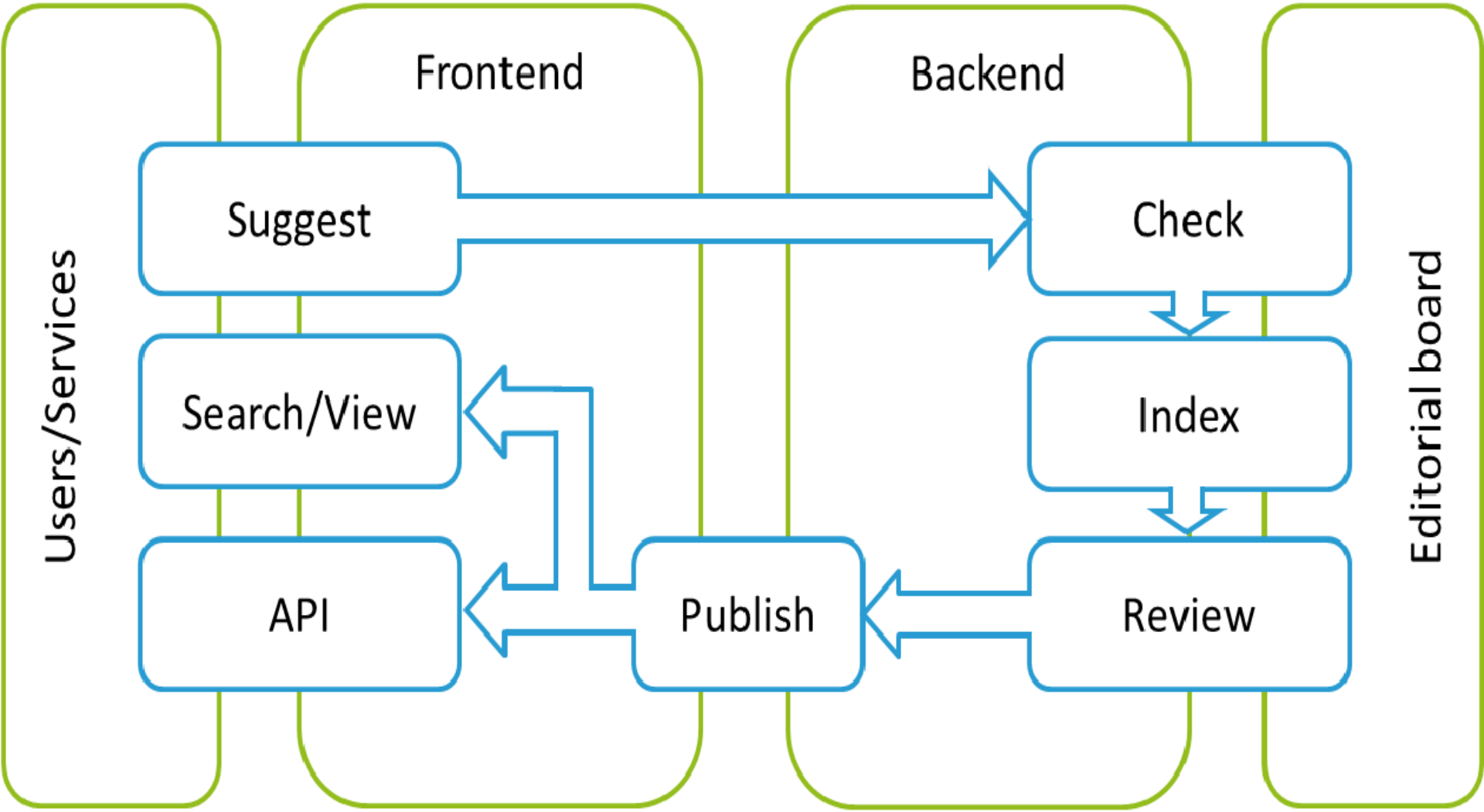


Figure 1: Research Data Repository Registration Workflow of re3data.org

RDRs IN INDIA

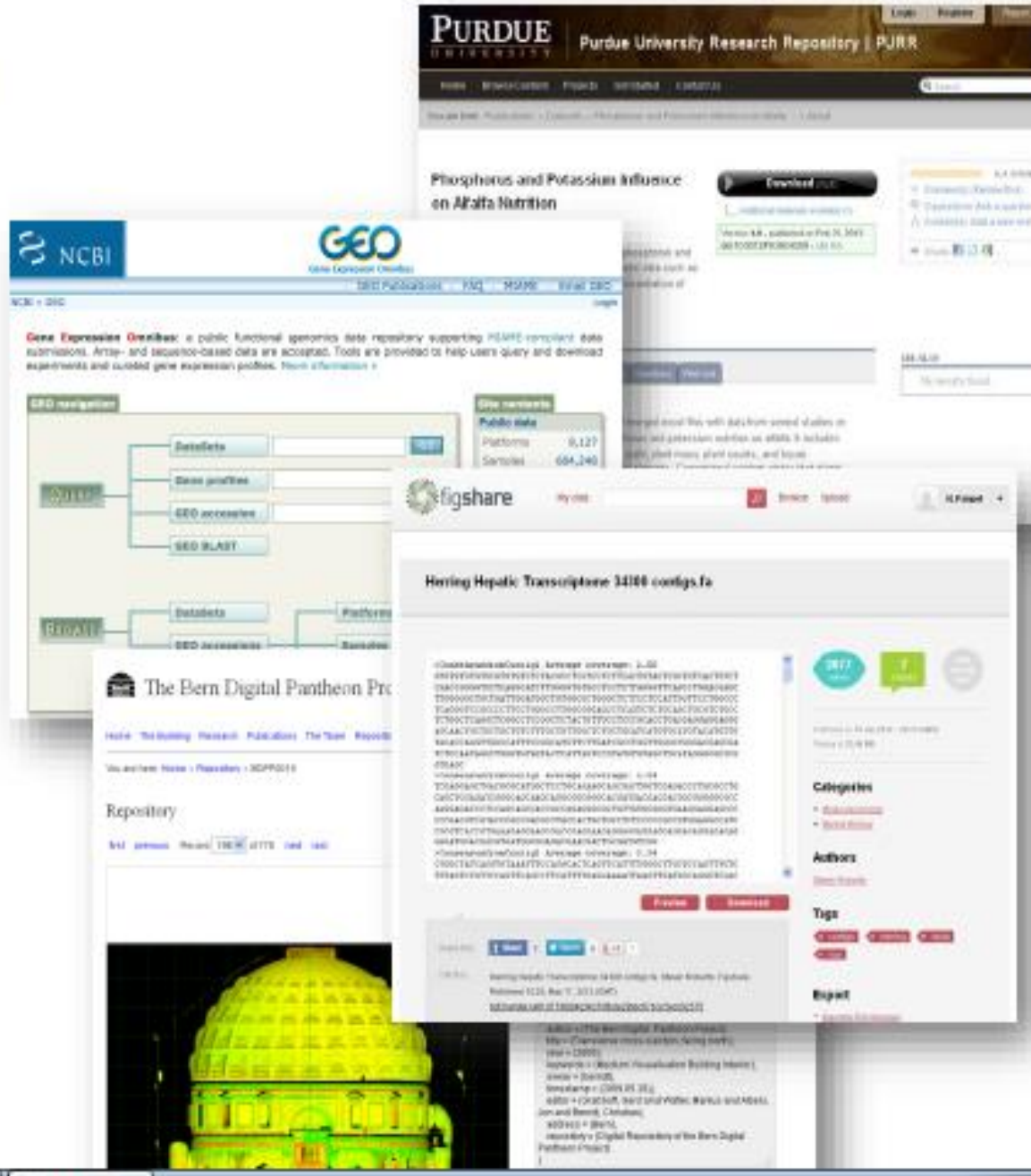
- **General information**
  - short description of the RDR, content types, keywords
- **Responsibilities**
  - institutions responsible for funding, content or technical issues
- **Policies**
  - policies of the RDR, incl. There URL
- **Legal aspects**
  - licenses of the database and datasets
- **Technical standards**
  - APIs, versioning of datasets, software of the RDR
- **Quality standards**
  - certificates, audit processes

Responsibilities of RDR

- The repository ensures that the deposited dataset is archived according to the FAIR principles to the best of its ability and resources.
- The repository preserves data collections for at least ten years after publication of the dataset policy for storage and management of research data.
- The repository shall, as far as possible, preserve the dataset unchanged in its original format, taking into account current technology and the costs of implementation.
- The repository has the right to modify the format and/or functionality of the dataset if this facilitates the digital sustainability, distribution, interoperability or re-use of the dataset.

RDR Typology

- Institutional
- Disciplinary
- Multidisciplinary
- Project



Functional Requirements for RDR

Metadata	• Support for different metadata schemas, support for domain-specific metadata, allowextended metadata for interoperability • Support data labeling by data owner, authorized persons or automatic metadata extraction tool • Provide metrics for metadata quality assessment • Support storing XML files • Provide metadata management tools • Support search engine indexing • Allow adding fields to collection schema
Identifiers	• Support assignments of PID and DOI • PID assignment for data management
Authentication and authorization	• Support integration with external systems • Provide single sign-on or support multiple authentication methods (Shibboleth, LDAP)

Quality = FAIR complete

- DataCite was queried to determine **where** researchers are sharing their research data
- Metadata **completeness analyzed using FAIR** recommendation for DataCite metadata
- to determine where researchers are sharing their research data and assessed the quality of the metadata.
- How are researchers making decisions about why and how to share research data?
- What is the cost to the institution to implement federally mandated public access to research data policies

METADATA ANALYSIS: TYPE

- Types of metadata
  - Descriptive
  - Structural
  - Administrative
  - Technical
  - Preservation
  - Access/Rights

DESCRIPTIVE METADATA

```
_
_vn 2.2ml
_pd 19970307
_mi M
_t1 CM000257 ANZ7371
_ml 01012081
_ma Goodwin, John Abbot,
_mu
_mt The Pilgrim republic;   an historical review of the colony of New Plymouth, with
sketches of the rise of other New England settlements, the history of Congregationalism,
and the creeds of the period.
_mo
_me
_ci Boston,
_pu Ticknor and Company; [etc., etc.]
_md 1888.
_mc
_mr
_ms New England -- History -- Colonial period, ca. 1600-1775.
_ms Puritans.
_t2 CM000257 ANZ7371 t2
_mn 712
_mp 000A 000B R001 R002 R003 R004 R005 R006 R007 R008 R009 R010 R011 R012 R013
    R014 R015 R016 R017 R018 R019 R020 R021 R022 R023 R024 R025 R026 R027 R028
    R029 R030 R031 R032 R033 R034 R035 R036 R037 R038 R039 R040 R041 R042 R043
    R044 R045 R046 0001 0002 0003 0004 0005 0006 0007 0008 0009 0010 0011 0012
    0013 0014 0015 0016 0017 0018 0019 0020 0021 0022 0023 0024 0025 0026 0027
    0028 0029 0030 0031 0032 0033 0034 0035 0036 0037 0038 0039 0040 0041 0042
    0043 0044 0045 0046 0047 0048 0049 0050 0051 0052 0053 0054 0055 0056 0057
    0058 0059 0060 0061 0062 0063 0064 0065 0066 0067 0068 0069 0070 0071 0072
    0073 0074 0075 0076 0077 0078 0079 0080 0081 0082 0083 0084 0085 0086 0087
    0088 0089 0090 0091 0092 0093 0094 0095 0096 0097 0098 0099 0100 0101 0102
    0103 0104 0105 0106 0107 0108 0109 0110 0111 0112 0113 0114 0115 0116 0117
    0118 0119 0120 0121 0122 0123 0124 0125 0126 0127 0128 0129 0130 0131 0132
    0133 0134 0135 0136 0137 0138 0139 0140 0141 0142 0143 0144 0145 0146 0147
    0148 0149 0150 0151 0152 0153 0154 0155 0156 0157 0158 0159 0160 0161 0162
```

STRUCTURAL METADATA

```
- <mets:structMap TYPE="mixed">
  - <mets:div LABEL="Proceedings of the Third Berkeley Symposium on Mathematical Statistics and
      Probability" TYPE="text" DMDID="DMD1">
    - <mets:div LABEL="Volume 1: Statistics" TYPE="volume" DMDID="DMD2">
      - <mets:div LABEL="Frontmatter" TYPE="frontmatter">
          <mets:fptr FILEID="FID1" />
        </mets:div>
      - <mets:div LABEL="Estimation of Least Squares and by Maximum Likelihood" TYPE="article"
          DMDID="DMD3">
          <mets:fptr FILEID="FID5" />
        </mets:div>
      - <mets:div LABEL="On a Use of the Mann-Whitney Statistic" TYPE="article" DMDID="DMD4">
          <mets:fptr FILEID="FID6" />
        </mets:div>
      - <mets:div LABEL="Index" TYPE="index">
          <mets:fptr FILEID="FID3" />
        </mets:div>
      </mets:div>
    - <mets:div LABEL="Volume 2: Contributions to probabilty theory, part 1" TYPE="volume"
        DMDID="DMD5">
      - <mets:div LABEL="Frontmatter" TYPE="frontmatter">
          <mets:fptr FILEID="FID2" />
        </mets:div>
      - <mets:div LABEL="On a Class of Probability Spaces" TYPE="article" DMDID="DMD6">
          <mets:fptr FILEID="FID7" />
        </mets:div>
      - <mets:div LABEL="Stationarity, Boundedness, Almost Periodicity of Random-valued
          Functions" TYPE="article" DMDID="DMD7">
          <mets:fptr FILEID="FID8" />
        </mets:div>
      - <mets:div LABEL="Index" TYPE="index">
          <mets:fptr FILEID="FID4" />
        </mets:div>
      </mets:div>
      <mets:div LABEL="Volume 2: Contributions to probability theory, part 2" TYPE="volume"
        DMDID="DMD8" />
    </mets:div>
```

TECHNICAL METADATA

```
<DeviceSource>digital still scanner</DeviceSource>
- <ScanningSystemCapture>
  - <ScanningSystemHardware>
      <ScannerManufacturer>Scitex</ScannerManufacturer>
    - <ScannerModel>
        <ScannerModelName>Leaf Volare</ScannerModelName>
        <ScannerModelNumber>100SX</ScannerModelNumber>
        <ScannerModelSerialNo>1122334455</ScannerModelSerialNo>
      </ScannerModel>
    </ScanningSystemHardware>
  - <ScanningSystemSoftware>
      <ScanningSoftware>Leaf</ScanningSoftware>
      <ScanningSoftwareVersionNo>4.0</ScanningSoftwareVersionNo>
    </ScanningSystemSoftware>
  - <ScannerCaptureSettings>
      <PixelSize>3</PixelSize>
    - <PhysScanResolution>
        <XphysScanResolution>0.307</XphysScanResolution>
        <YphysScanResolution>0.435</YphysScanResolution>
      </PhysScanResolution>
    </ScannerCaptureSettings>
  </ScanningSystemCapture>
- <DigitalCameraCapture>
    <DigitalCameraManufacturer>PhaseOne</DigitalCameraManufacturer>
    <DigitalCameraModel>H_20</DigitalCameraModel>
  </DigitalCameraCapture>
- <CameraCaptureSettings>
    <FNumber>5</FNumber>
    <ExposureTime>.5</ExposureTime>
    <Brightness>2.2</Brightness>
    <ExposureBias>3.5</ExposureBias>
    <SubjectDistance min="4.9" max="5.3">5</SubjectDistance>
    <MeteringMode>Average</MeteringMode>
    <SceneIlluminant>D75 Illuminant</SceneIlluminant>
    <ColorTemp>20</ColorTemp>
    <FocalLength>3</FocalLength>
    <Flash>No</Flash>
```

DataCite

Find what you're looking for by searching millions of records with extensive, reliable metadata.

Share your data and reuse the data of others to create the highest impact in the research community.

Cite your research sources with confidence, and receive proper credit when your work is reused.

Connect your research: publications, datasets, software, authors, institutions, and funding data all in one place.

DataCite is a leading global non-p
organisation that provides persist
identifiers (DOIs) for research dat

Our goal is to help the research co
munity locate, identify, and cite
research data with confidence.

More than 40 members and 80
data centers assign DataCite DO

We invite members from all types
organisations: data centers, publish
libraries, funders, and others to join
support our data sharing mission

http://www.datacite.

Recommended and Optional Metadata Properties (v. 4.2)

ID	Property	Obligation
6	Subject (with scheme sub-property)	R
7	Contributor (with optional given name, family name, name identifier and affiliation sub-properties)	R
8	Date (with type sub-property)	R
9	Language	O
11	AlternateIdentifier (with type sub-property)	O
12	RelatedIdentifier (with type and relation type sub-properties)	R
13	Size	O
14	Format	O
15	Version	O
16	Rights	O
17	Description (with type sub-property)	R
18	GeoLocation (with point, box and polygon sub-properties)	R
19	FundingReference (with name, identifier, and award related sub-properties)	O

Datacite Metadata Elements

Element name	Obligation level	Mandatory type
Identifier	Mandatory	Descriptive
Creator	Mandatory	Descriptive
Title	Mandatory	Descriptive
Publisher	Mandatory	Descriptive
Publication year	Mandatory	Descriptive
Resource type	Mandatory	Technical
Subject	Recommended	Descriptive
Contributor/s	Recommended	Descriptive
Related identifier	Recommended	Structural
Date	Recommended	Descriptive
Description	Recommended	Descriptive
Geolocation	Recommended	Descriptive
Language optional	Optional	Descriptive
Alternate identifier	Optional	Structural
Size	Optional	Technical
Format	Optional	Technical

Dublin Core, expressed in HTML meta tags

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE html
    PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
     "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="en-US" xml:lang="en-US">
<head>
<title>Arithmetic</title>
<meta name="ROBOTS" content="INDEX, FOLLOW" />
<meta name="DC:title" content="Arithmetic" />
<meta name="DC:creator" content="Sandburg, Carl, 1878-1967" />
<meta name="DC:creator" content="Rand, Ted, ill." />
<meta name="DC:publisher" content="San Diego :Harcourt Brace Jovanovich" />
<meta name="DC:date" content="c1993" />
<meta name="DC:language" content="eng" />
<meta name="DC:description" content="A poem about numbers and their characteristics.
Features anamorphic, or distorted, drawings which can be restored to normal by viewing from
a particular angle or by viewing the image's reflection in the provided Mylar cone." />
<meta name="DC:description" content="One Mylar sheet included in pocket." />
<meta name="DC:subject" content="Arithmetic" />
<meta name="DC:subject" content="Children's poetry, American" />
<meta name="DC:subject" content="Arithmetic" />
<meta name="DC:subject" content="American poetry" />
<meta name="DC:subject" content="Visual perception" />

<script type="text/javascript">//<![CDATA[
        function preload_image(args)
        {
        var myimages=new Array();
        var i = 1;
        args++;
        while (i != args)
        {
        myimages[i] = new Image();
        myimages[i].src = preload_image.arguments[i];
        i++;
        }
        }
```

METADATA QUALITY

- Metadata quality refers to “the degree to which the metadata in question perform the core bibliographic functions of discovery, use, provenance, currency, authentication, and administration.” **Park (2009)**
- When evaluating metadata quality, the conformity to a set of requirements is determined.
- A list core criteria used for assessing metadata quality : completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. **Bruce and Hillmann (2004).**
- Metadata quality criteria can either be applied to individual metadata elements, to metadata records, or to entire metadata collections **Zeng and Qin (2022)**.

CONCEPTUAL FRAMEWORK

- ***General section*** which includes the types ,formats and granularity of the metadata to assess provision. Under this section four metadata core criteria are available such as completeness, comprehensiveness ,appropriateness and accessibility.
  - *Completeness* consists of the use of individual metadata elements are described completely i.e. the number of metadata elements used in a metadata record in relation to the number of a metadata elements available. The individual metadata elements used here indicates how frequency a metadata element is used in the sample of metadata records.
  - *Comprehensiveness* of metadata description deals with the use of element description i.e. the number and combined character length of descriptions in a metadata record.
  - *Accessibility* consists of the metadata used can be easily accessed without any difficulties.

CONCEPTUAL FRAMEWORK CONTD....

- ***Tools and technique*** which deals with the structure, application of semantic web technologies ,indexing and use of terminologies to assess the metadata. In this section the main criteria are accuracy, discoverability ,interoperability ,extendibility etc.
- ***Usability*** section refers to the presence in repositories, application of semantic mappings, metadata standards and cross-walks provision. It consists of *Conformance to expectation* i.e. the metadata is described in such a way to meet the expectations of the user. Another criteria is Logical consistency and coherence which means the metadata elements are consistent with standard definitions and description should be coherent across collection.
- ***Management and Curation*** section deals with two parts. First is the creation and version of the metadata being used and second the creation and version information used itself i.e. meta-metadata of the quality assessment itself .The main criteria in this section are timeliness, versionability and meta-metadata.

Metadata quality criteria	Description
Completeness	Use of individual metadata elements are described completely i.e. the number of metadata elements used in a metadata record in relation to the number of a metadata elements available. The individual metadata elements used here indicates how frequency a metadata element is used in the sample of metadata records
Accessibility	Extent to which metadata can easily accessed without any difficulties.
Comprehensiveness	Use of element description i.e. the number and combined character length of descriptions in a metadata record.
Appropriateness	Metadata and data documentation appropriately describe data
Accuracy	Metadata elements are described correctly.
Discoverability	How the metadata are easily found.
Conformance to expectation	Metadata is described in such a way to meet the expectations of the user
Logical consistency and coherence	Metadata elements are homogeneous and constant. They are consistent with standard definitions and description should be coherent across collection.
Open data licence	Data are assigned with an open licence
Reuse potential	The dataset is analyzed by others in future.
Interoperability	Extent to which metadata can be exchanged and used without any problem
Timeliness	Metadata is current having temporal information.
Versionability	Extent to which a new version may be easily created.
Meta-metadata	Metadata about the metadata.

Metadata Accuracy

```
<DC_record>
  <creator>Mitchell, William J.</creator>
  <creator>Stevenson, Daniel C.</creator>
  <creator>Schoonover, Regina</creator>
  <title>Urbanowski, Frank</title>
  <subject>City of Bits: Space, Place, and the Infobahn</subject>
  <subject>Electronically mediated environments</subject>
  <subject>Cyberspace</subject>
  <type>Urbanism</type>
  <format>Text</format>
  <date>text/html</date>
  <identifier>1995</identifier>
  <language>http://press.opt.edu/CityOfBits.html</language>
</DC_record>
```

Metadata Consistency

- DC records with a <format (e.g., <dc:date>YYYY-MMdc:date> element
- Most formatted in full W3C-DTF -DD</dc:date>),
- except for:

  <dc:date>2000-08</dc:date>

  <dc:date>1996</dc:date>

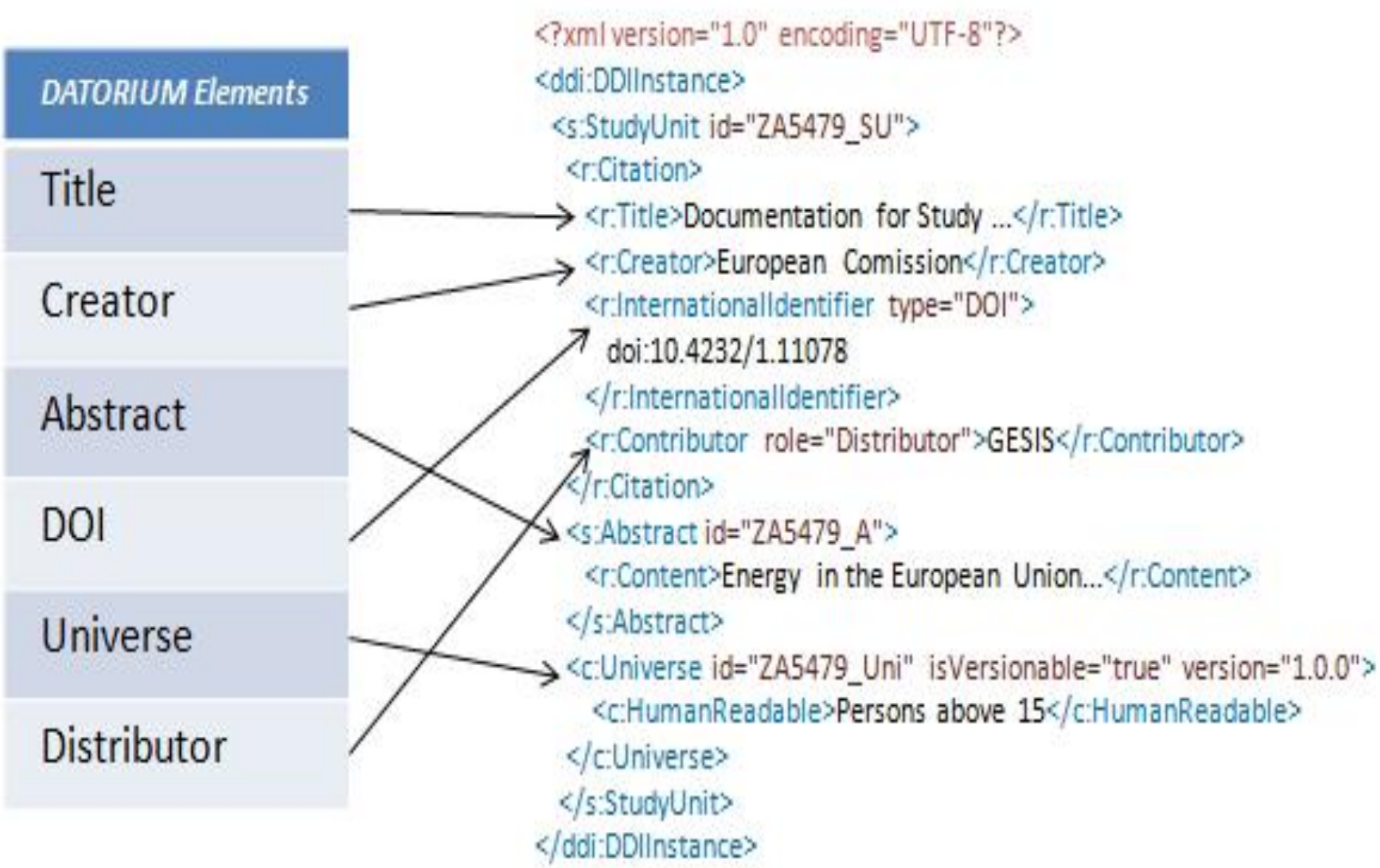
  <dc:date>July 5, 2001</dc:date>

  <dc:date>2000 Revision</dc:date>

  <dc:date>July 19, 1996</dc:date>

  <dc:date>2001.06.04</dc:date>

METADATA MAP



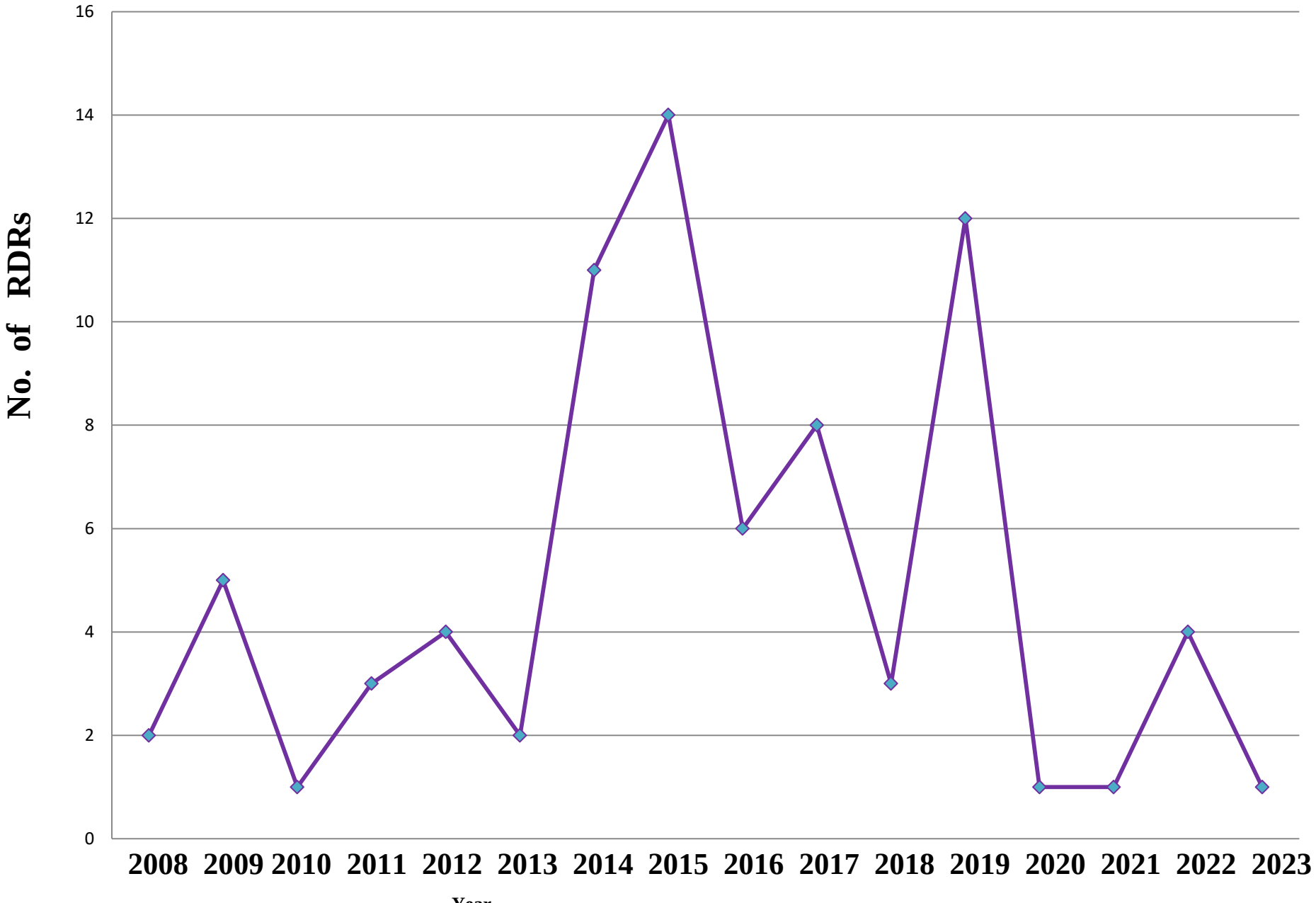
OBJECTIVES OF THE STUDY

- To identify Indian research data repositories Indexed in the re3data.org
- To trace year wise development of research data repositories (RDRs) in India
- To discover standards and specification being used in Indian research data repositories
- To assess software tools used to develop Indian research data repositories
- To understand the author identification system followed in managing data in research data repositories
- To examine various aspects of data access policies, persistent identifiers and application programming interfaces used in Indian RDRs.
- To know the metadata standards and schemas used in RDRs.
- To identify different criteria of metadata quality for assessment.

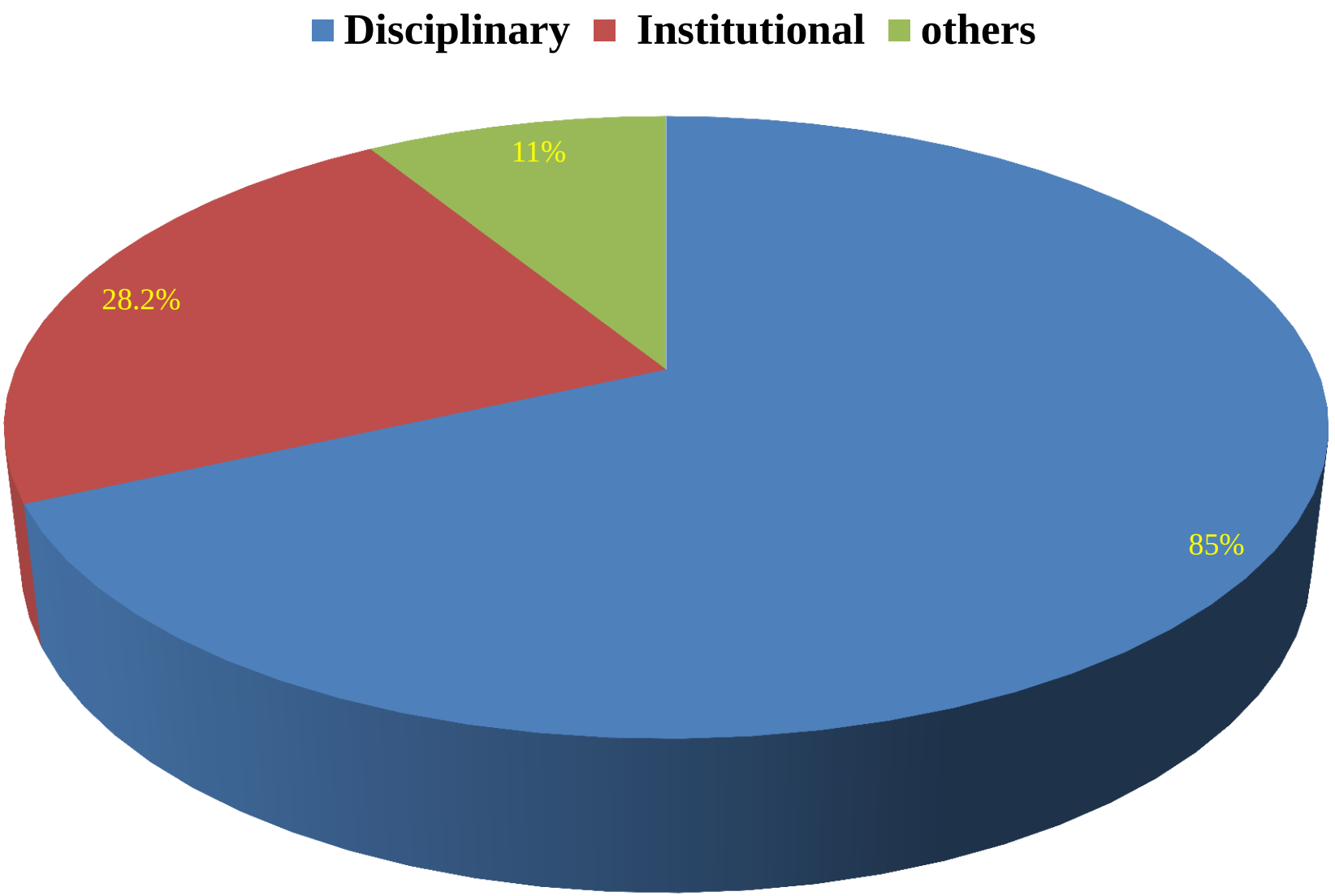
METHODOLOGY

- A survey method was conducted on the Research Data Repositories in India indexed in Registry of Research Data Repositories (r3data.org)
- The questionnaire received from 45 nos. of research data repositories all over India was 286 i.e. 81% .
- The software used for data analysis in the study - R Software

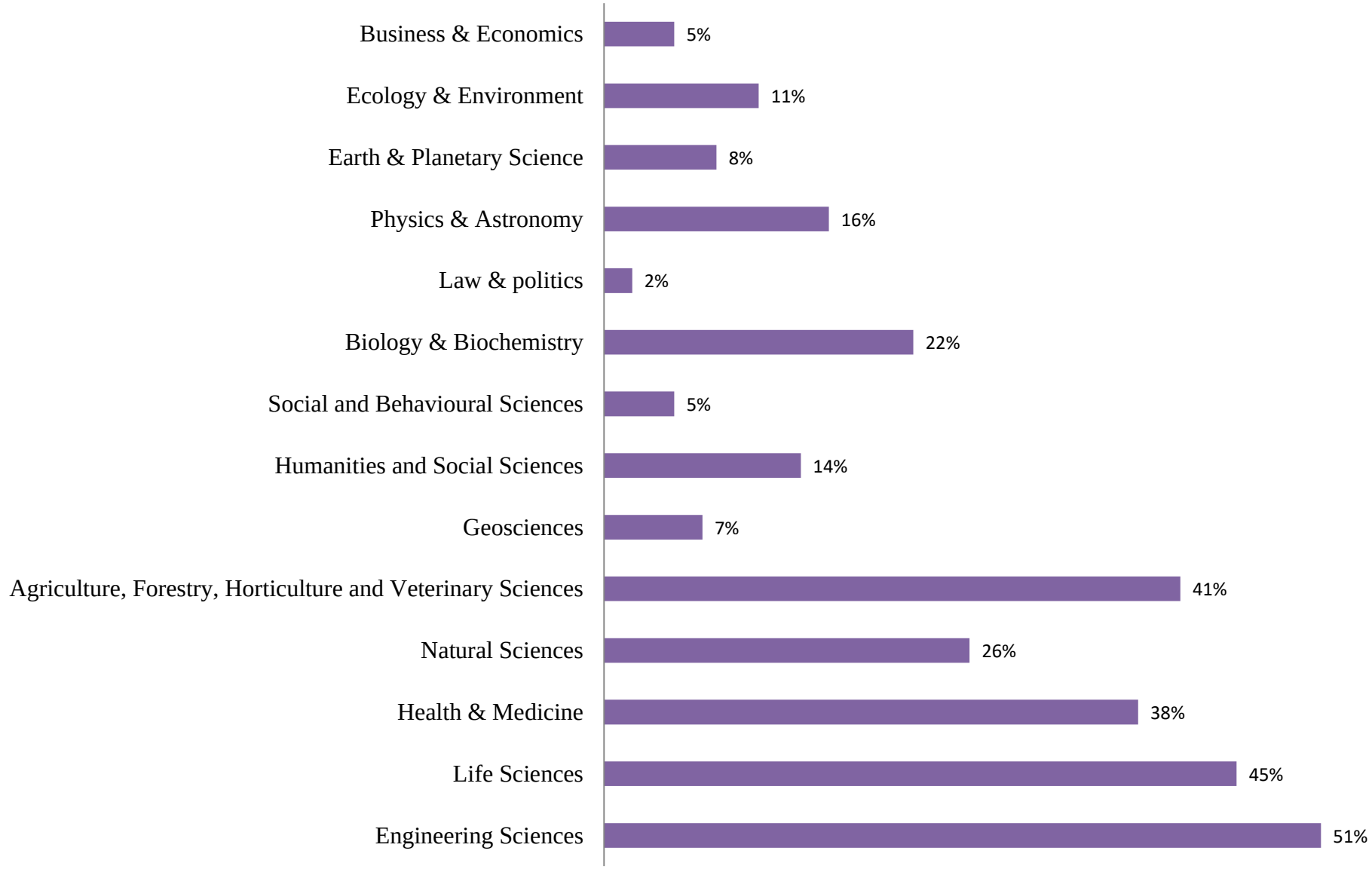
Year wise growth of research data repositories in India



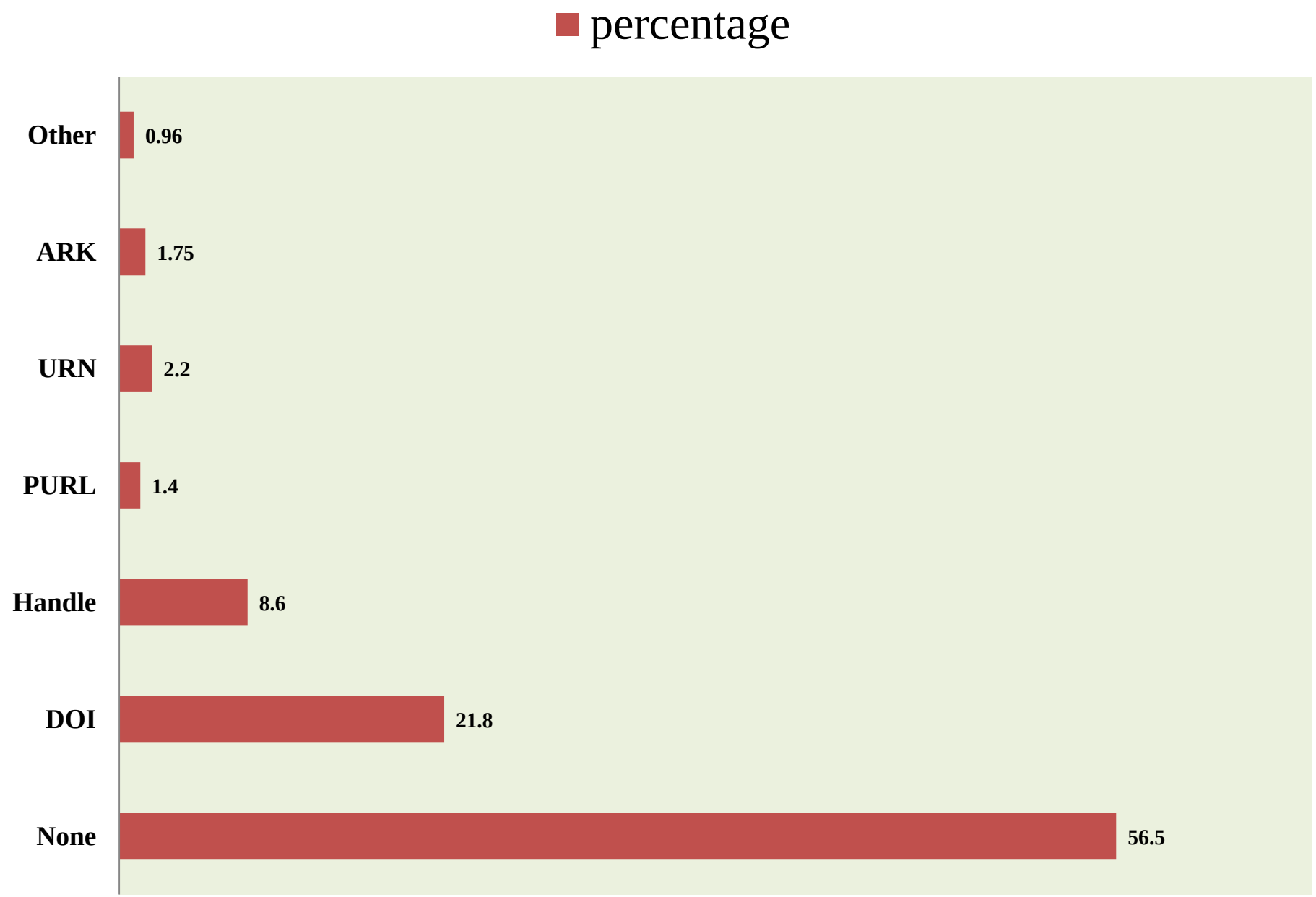
Type of research data repositories indexed in r3data.org(n=51)



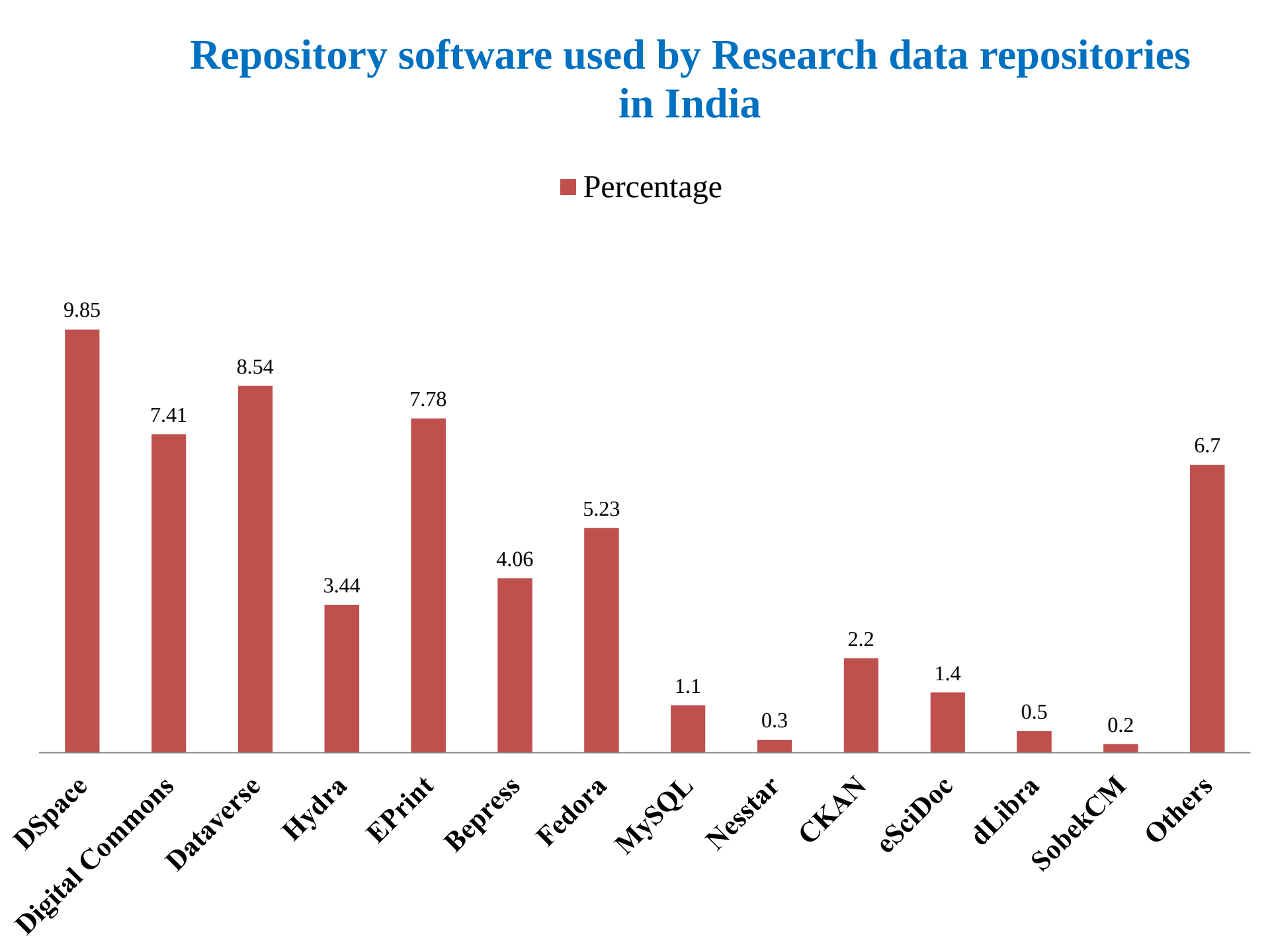
Distribution of research data repositories by subject coverage in India



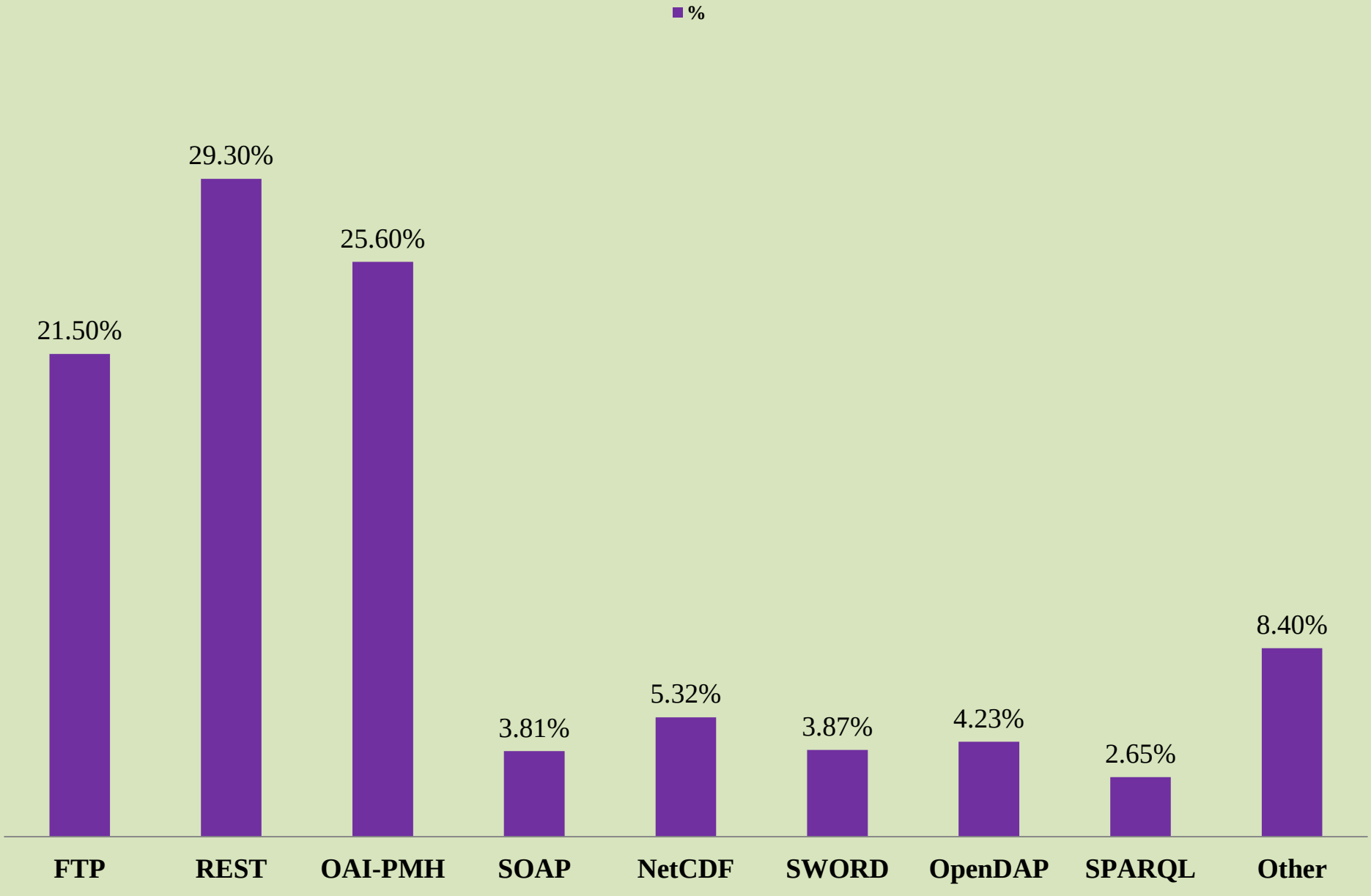
Persistent identifier systems used by research data repositories in India



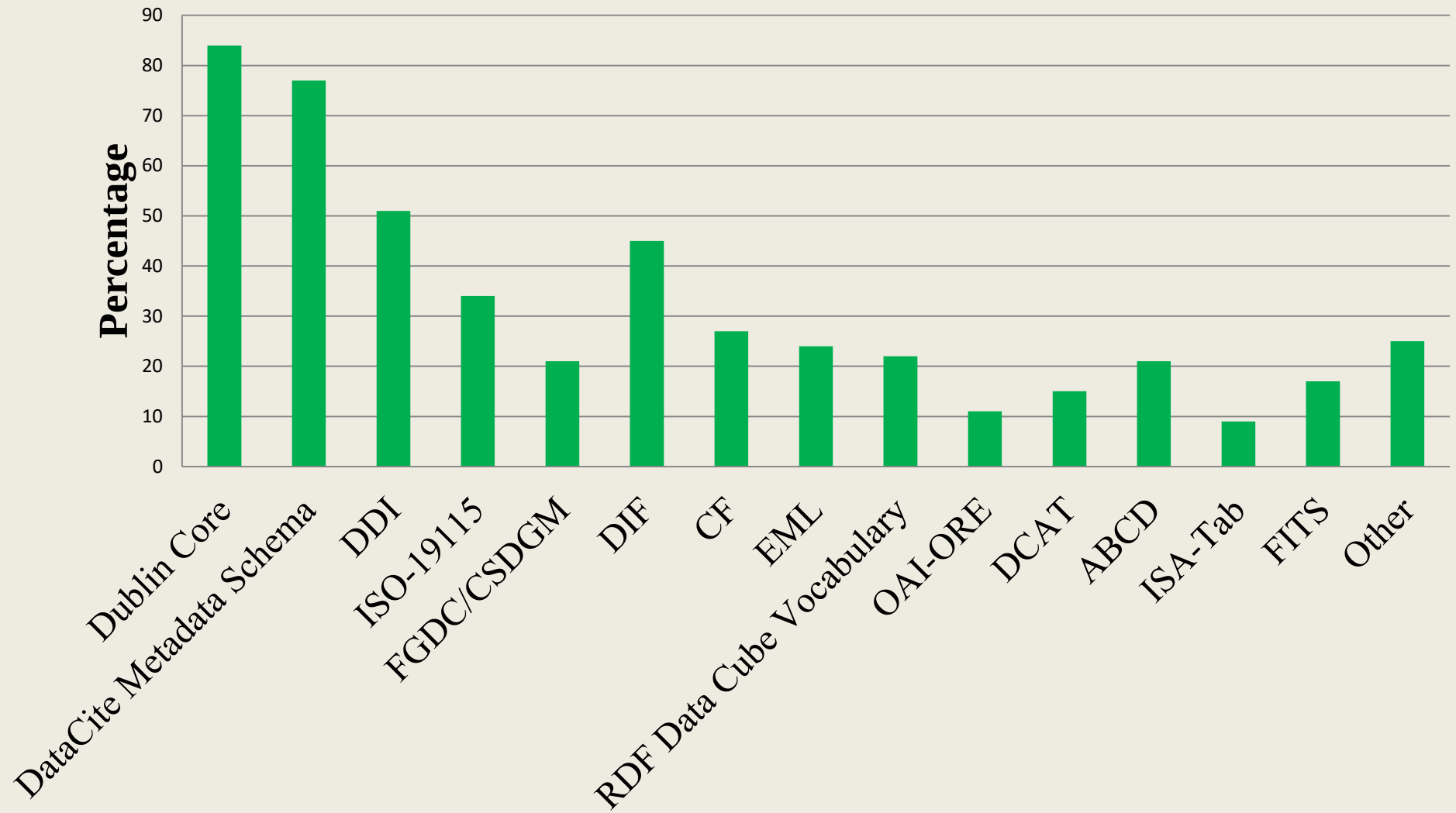
Repository software used by Research data repositories in India



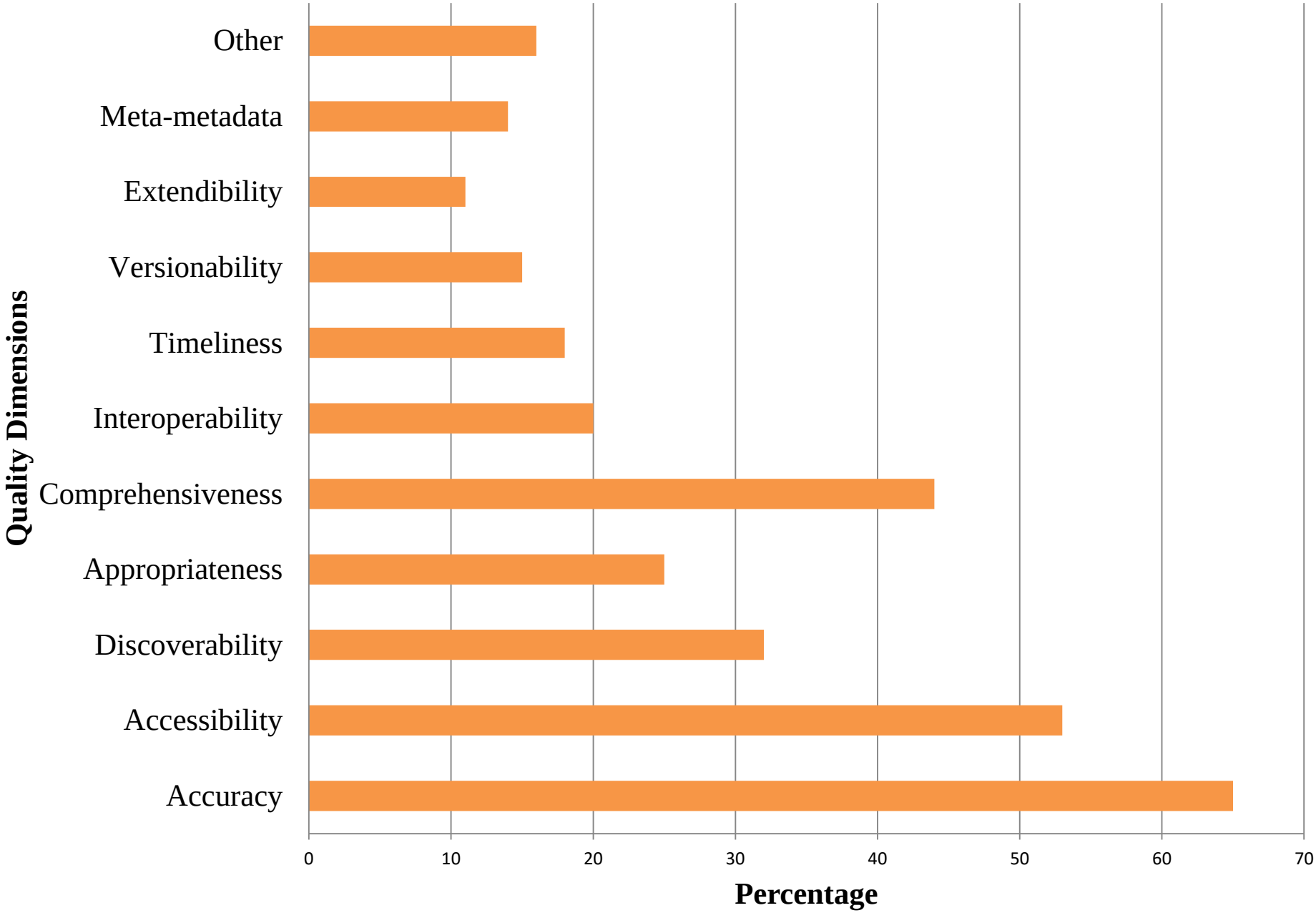
Distribution of Research Data Repositories by API System



Metadata Standards used in RDRs



Metadata Quality Criteria used in RDRs



FINDINGS

- Mandatory elements are used most frequently, followed by recommended and optional elements of individual metadata elements,. More than half of all metadata elements are used in less than 10 % of metadata records. With the exception of related identifiers, persistent identifiers are rarely used. The average descriptions has 597.3 characters. On average, 27.7 elements are used in metadata records, which corresponds to 11.7 % of the elements available. The homogeneity of metadata records varies considerably between repositories, on average, 51.6 % of metadata records use the same common set of metadata elements.
- The analysis revealed statistically significant differences across repositories of varying type and certification status in the use of individual metadata elements, the comprehensiveness of descriptions, and the completeness of metadata records.
- Completeness of metadata records vary across repositories, which could be an indicator for distinct metadata practices at individual research data repositories, but is likely also skewed by using a generic metadata schema for describing diverse datasets. Within repositories, metadata descriptions are relatively homogenous, suggesting that repositories have developed consistent practices for describing data.

CONCLUSION

- In order to identify such metadata, we then analysed, for each repository, the metadata requested at submission time and the metadata exposed at visualisation time, i.e., the metadata returned when a repository user access the dataset landing page.
- This paper presents a first systematic analysis of metadata quality for research data and the influence of repository characteristics on metadata quality. It discusses difficulties of using a generic metadata schema for describing diverse research data.
- The results show that some repositories appear to have established successful metadata practices and workflows, but some metadata elements remain underused. There is evidence of repository type and certification status affecting metadata quality, but more research is needed to identify specific factors.

THANK YOU

?