Standard Operating Procedure (SOP) for Digitization of Archival Records

National Archives of India
Janpath, New Delhi-110001
2024

CONTENTS

1. Background

The need for rapid digitization of archival records was discussed in the 46th meeting of the National Committee of Archivists(NCA) held at Bikaner on 4-5 September, 2023. However, it emerged during discussions that different standardsare being followed by various archives for digitization of records. There is no uniformity in the meta data formats and the storage methods being used. It was found in some cases that records had not been hosted on any public portal after digitization. All the members of NCA unanimously agreed that time bound plans should be drawn up for digitization of archives, and this work should be completed soon, preferably within the next five years. It was also decided that there is a need for a uniform standard operating procedure to be followed for digitization of records. A sub- committee comprising Heads of State Archives of Delhi and Gujarat was constituted to propose an SOP in this regard.

The committee duly submitted its report in March 2024, and this report was discussed during the 47th meeting of the National Committee of Archivists held at Srinagar on 18th-19th March2024. After detailed discussions, it was decided that the National Archives of India (NAI) may come up with a draft of the SOP to be followed for digitization, utilizing the report submitted by NCA Sub-committee as the base document. This SOP may be circulated to all members of NCA for comments and then finalized after suitably incorporating the comments.

The National Archives of India framed comprehensive Guidelines for Digitization for records with the objective of laying down SOP for handling and digitization of archival records, and circulated to all the State/UT Archives and Archives in India Online (AIO) members for comments/views through a DO letter on 4 June 2024 for finalization by the National Committee of Archivists. The dateline for submission of comments/views was marked on 30 June 2024. In response, a total of five comments from (1) Department of Delhi Archives, (2) Kerala State Archives (3) Anandashram Sanstha, Pune, (4) Bihar State Archives and (5) Kerala State Archive were received. The National Archives of India has incorporated the comments with the justifications whose details are mentioned on page No. 22-25.

This document proposes the standards for creating archival quality digital still images of archival materials such as manuscripts, printed books, maps, photographs, etc. It covers the major points which should be considered before and after digitization, including image quality, file formats for storage and accessibility standards for scanned images and born digital records.

2.Image Capturing (Scanning):

1. Resolution: The records should be digitized at minimum of 300-600 dpi true optical resolution. They should be scanned in the 24 bit colour mode to capture original information truthfully. Specifically, text documents are to be scanned at 300 dpi (in case of legibility issues 400 dpi) and images/ photographs/ treaties/ manuscripts are to be scanned at 600 dpi.

2. Scanner: should use Face up Scanning technology with pixel type moving linear CCD sensor overhead scanner or CMOS overhead scanner.Scanner should use cold light with uniform illumination and such low intensity oflight during scanningthat it does not damage the fibers of ageing pages and prevents any harm to the original records. The scanner must be capable of tightly controlling non-linearities and quantization noise, in order to handle records of variable density and achieve optimized results for all types of records. Cradle type scanners should be used, so that there is no need to unnecessarily unstitch documents for scanning.

3. However, if stitching renders some of the content unreadable, then unstitching must be carried out to scan the complete information.

4. Coloured charts, diagrams, photos, illustration etc. are to be scanned separately wherever applicable.

5. Original scanned records should be captured in TIFF v6.0 Format with LZW compression (ISO 12230-2:2001) or later.

6. All the digital images should have a checksum for the authenticity of digital image.

7. Once scanned along with checksum, no changes should ever be made in the TIFF file. It can only be used as a base for further image processing, without undergoing any editing itself.

3.Image processing and cleaning (Image Enhancement):

1. The TIFF images will have to be further enhanced to increase the legibility of the text and overall visual appeal of the document without altering the authenticity feel of the original source.

2. Image enhancement activities should be carried out forcropping, bleed through removal, black border removal, curvature correction, light equalization, brightness and contrast enhancement, de-noising, de-skew and de-speckling, cleaning, sharpness, focus enhancement, background removal for text areas without violating content in picture zones and without altering the original capture dpi.

3. The removal of digital noise will include removal of worm-marks and stain-marks to the maximum possible extent, while keeping the colour information intact. This component of digital restoration will aim at attaining a relatively high level of noise-free state of the digital images.

4. Proper editing has to be carried out to straighten the orientation of the content matter if there is any disorientation existing in the pages.

5. 100% manual operator assisted quality check should be performed on every image to ensure that there are no missing or dropped pages and that images are processed to their optimum levels.A detailed quality checklist has to be certified for every page being processed.

6. Nuances in the document such as notes, remarks, and pencil marks are of historical value and they must be preserved in such a way that may be referred in lieu of the original document.

7. Image processing priorities should be legibility, aesthetics and file size in that order.

8. A thumbnail image should also be generatedfor each document.

9. The enhanced images will have to be delivered in the JPEG v1.02 Format (ISO DIS 10918-1/ ISO/IEC10918-5:2013) or later.

4.Optical Character Recognition/ Handwriting Recognition, Translation into English and AI based auto tagging.

1. Optical Character Recognition should be carried out on the JPEG files.

2. Over 95% OCR accuracy is desired to be achieved for digitized documents with printed/ typewritten text.

3. OCR should support recognition of multiple languages on the same page.

4. Text should be Unicode, supporting all major Indian languages.

5. In case of handwritten documents, handwriting recognition should be carried out. At least 50% accuracy has to be ensured in HWR and multiple language support should be supported.

6. For facilitating easy search operations, it would be desirable that all texts which are not already in English, should be translated into English.

7. AI based auto tagging should be carried out to generate keywords for all scanned documents. These keywords will facilitate more meaningful search operations than merely looking at the subject lines of documents.

5. Conversion to PDF/A Format

1. The JPEG files will be converted after the OCR process to a searchable PDF/A format with the text layer below the original image.

2. Searchable PDF/A should contain Watermark showing the source archive's name on all pages.The watermark should be secured from modification and extraction of the text layer and 256 bit AES encrypted.

3. The OCR/ HWR text extracted from each page, auto tags generated and metadata should form part of the PDF/A file. It would be desirable to include the quality check certificate also within the PDF/A file.

4. The searchable PDF/A file should be compressed over 90% without altering the original capture dpi, retaining of the page dimensions and without losing the legibility of the document.

5. The PDF file will be delivered in the PDF/A Format (ISO 19005:1/ISO19005-1:2005)or later.

6. The archive should use standardized metadata format and appropriate technology to extract text from the images and provide a separate data base, so that the same can be integrated in the NAI search portal i.e. www.abhilekh-patal.in.

6. Subject Metadata and Captioning

1. Metadata of the all the digitized files describing the content of scanned document, in searchable format, should be assigned as per the standard format enclosed at **Annexure 1.**

2. Metadata has to be provided in .CSV file and XML file format with every batch of digital records for online interoperability.

3. The metadata indexing has to be done with the right spelling, punctuation, grammar and information.

7. Unique record identifiers-

1. Preferably, each record should be assigned a unique bar code, which should be physically pasted onto the record using archive grade adhesive. This bar code should be incorporated in the record management database also to facilitate location of records.

2. Bar coding should invariably form a part of the digitization process.

3. In case some documents have already been digitized and bar codes have not been assigned, it should be ensured that all such documents are also assigned bar codes.

4. Some archives are already using unique record identifiers, which are numeric or alphanumeric strings that are associated with a single entity i.e. an e-record within a given system. The unique record identifier serves as a “filename” for an e-record. The e-record producing organization should define a set of rules or conventions for generating the unique record identifiers.

5. Such unique record identifiers should also be linked to the bar codes assigned, so that bar code scanning can be conveniently used to locate records.

8.Quality control checklist

Every digitized image prepared from record should be checked for quality and compliance to the following minimum checklist of quality control of scanned files must be ensured:

- File name is correct
- File format is correct
- Bit depth is correct See: file, properties, details
- Image is correct size/resolution in long dimension
- Image is not skewed or off centered
- Image has clean edges, clear contrast, and legible text
- No broken figures (illustrations, maps, etc.)
- No more patterns (wavy lines or swirls, usually found in areas where there are repeated patterns)
- No presence of digital artifacts (such as very regular, straight lines across picture)
- No pixelation (individual pixels are apparent to the naked eye)
- Not too light or too dark
- No loss of detail in highlight or shadows
- OCR/ HWR accuracy as specified
- Metadata for the respective record/image/manuscript/paintings, etc. should be checked for accuracy and completeness
- Spell check of the captured metadata

The digital images should be prepared in the below mentioned standards of TIFF, JPEG and PDF/A format along with Metadata in XML or CSV format. Checksum with the pages collated as per the original sequence and page integrity should be maintained. The Quality Check file in CSV format and a Thumbnail image file should also be created for each document.

Technical standards of TIFF, JPEG & PDF Images**:

i. TIFF v6.0 LZW Compression (ISO 12230-2:2001)
ii. JPEG v1.02 (ISO DIS 10918-1 / ISO/IEC 10918-5:2013)

iii. Searchable PDF / A (ISO 19005:1 / ISO 19005-1:2005)

**

- Any updation in above mentioned ISO standards will be duly considered.
- All files should be able to pass JHOVE format validation as valid and well-formed.

9. Access to digitized files/records

1. All the digitized files/records should be made accessible from adedicated portal linkedwith NAI search portal i.ewww.abhilekh-patal.in. Any visitor should be able to perform a search on this portal without having to log in. There may be a requirement of free registration on the portal to view the contents of a file, and at most a nominal charge may be prescribed for downloading of content.

2. If there is any problem in creating a web portal for viewing of documents, the archive concerned may explore the option of utilizing the NAI portal as a distinct sub-chapter for this purpose, on payment of minimal charges that will be incurred on hiring of additional cloud space.

10. Long Term Storage

1. All the digital images in three formats(TIFF, JPEG and PDF/A) along with metadata should be stored in long term storage with Disaster Recovery provisions.

2. For this purpose, one copy of the data may be kept in archival (cold) cloud storage, which is possible at low costs since the frequency of access will be extremely low. Users will normally access the PDF/A image only from the web portal, for which a hot cloud storage will be separately contracted.

3. The cloud service provider must be asked to ensure that the data provided to them will be available at all times.

4. NAI has avoided setting up of their own Data Centre(DC) and Recovery Centre(RC), since proper maintenance of such facilities does not fall within its core competence. It is felt that professionally managed cloud services will be more reliable.

5. It would be preferable to go for MEITY empanelled vendors in the interest of keeping data secure. Data must be hosted on servers based in India.

6. The second copy of data should be within physical possession of the archive, in its own premises. It should have an air gap, i.e., it should not be accessible over the internet or by any other remote means, in order to rule out the possibility of data corruption/ hacking.

7. Looking at the state of technology available today, such storage could most conveniently be in optical or magnetic storage. NAI has opted for storage in LTO cartridges, protected by an electro-magnetic radiation proof vault within its own premises.

8. Proper arrangement for migration of data to new optical/ magnetic media as per media longevity specifications must be made.

11. Modalities of digitization

1. A timebound plan for digitizing all documents in an archive typically requires a large number of scanners and manpower for a limited period of time.

2. Parallelly, a large number of documents may need to be repaired. This is a very labour-intensive task for which an even larger number of manpower is required. This requirement is also for a relatively short period of time only.

3. It is generally not possible, nor is it advisable, to create a large number of posts for a limited period of time.

4. The best approach for accelerating digitization work would be to select a vendor through competitive bidding process. While doing so, a Quality cum Cost Based Selection instead of purely cost based selection would generally provide better results[1].

5. Separate vendors can be selected for repair of documents and for checking the quality of scanned documents, if necessary.

[1] NAI has already engaged a Digitization Vendor for scanning 600,000 pages per day; a Quality Check Vendor to check the physical documents being returned and the quality of digital images; a Repair Vendor to repair/ laminate 45,000 sheets per day; and is in the process of finalizing an IT Solutions Vendor for long term storage of data. Any archive may obtain the bid documents for guidance purposes, should they wish to do so.

APPENDICES:

Annexure- 1: Standard format

#	Meta Fields	Meta value	Example
1.	Bundle Barcode	14Digit Unique Code (Alpha Numeric)	NAIDLB0000200 (for bundles /volumes)
2.	File Barcode	14Digit Unique Code (Alpha Numeric)	NAIDLF0000200 (for files)
3.	Location	Alpha-Numeric	Eg- Repositories in Delhi headquarter i.e R1, R2,R3,R4, PA, Cartography, OR, Lahore shed, Bhopal, Jaipur , Pondicherry, Bhubaneswar etc.
4.	Location Code	14Digit Unique Code (Numeric)	Inclusive of Room Master Numbers (Three Digit), Compactor (One Digit), Pillar Number or Compactor Number (Three Digit), Rack Number (Three Digit), Shelf Number (Two Digit), Location of Bundle or Volume (Two Digit). Eg-003-1-005-001-12-01 Eg-003-1-005-001-12-02
5.	Identifier	Unique no. (Alpha Numeric)	Eg- PR No. LiB. No, PA No. , microfilm no.
6.	Category	Alphabetical.	Public records, Private Archives, cartography , oriental records, photographs, microfilm, oral Archives, etc
7.	Status of digitization	Alphabetical.	Digitized Document/ Non- Digitized document.
8.	Document Title	Alpha-numeric.	Title of the documents.
9.	Keywords	Alpha-numeric.	Keywords to search a file
10.	Scale	Numeric.	Basically used in Cartographic records.
11.	Ministry/	Alpha-	Eg-Ministry of Defence, Home

	Department/ Residency	numeric.	Public department, Rajputana Residency.
12.	Branch	Alpha-numeric.	World War-II
13.	From Year - Date(mm-dd-yyyy)- To year Date(mm-dd-yyyy)	Numeric.	Eg- 1857-1912 1943-1946
14.	Geographical references	Alphabetical	Eg- Prant/Thana/Tehsil/District/Block No./ Panchayat/ Taluks, administrative units etc. It depends according to series or the Origin of records, or place where it belongs to.
15.	File No./Reference No./Sheet no./Folio no.	Alpha-Numeric	Eg- 601/4297/WD or the daftar numbers, reference numbers can be filled in this column.
16.	Part Number	Alpha-Numeric	Eg- Volume-I/Part -I
17.	Physical condition of the record.	Alpha-Numeric.	Eg- delicate, loose sheets, preserved etc.
18.	Document last date of repair	Alphabetical.	Month/ Year - Physical/ Microfilm/Digitized
19.	Series	Alpha-Numeric	Eg- War Diaries, QMG etc.
20.	Source organization	Alphabetical.	E.g- National Archives of India
21.	Record Contains	Alpha-Numeric	If any Maps, Photographs, Albums, Seals, coins , illustrations etc.
22.	Language	Alphabetical	English/Urdu/Persian/Hindi/Gujarati/ Tamil Etc
23.	Microfilm Roll Number	Alpha-Numeric	Microfilm Rolls
24.	Relation	Alpha-Numeric	Content relates to other resources.
25.	Copyrights	Alpha-Numeric	Copyright info.
26.	Call Number	Alpha - Numeric	Subject related number provided in the libraries.
27.	Publisher	Alpha Numeric	Oriental Records, Library books/PA.
28.	Subject	Alpha-	Topic covered.

		numeric.	
29.	Creator	Alpha-numeric.	Organization/Author name of the book.
30.	Accession Number	Alpha-numeric.	In OR/ PA/Library
31.	Year of publication/ (YYYY-MM-DD)	Numeric	In OR/Library/PA
32.	ERA	Alpha-numeric.	Eg-URLs, DOIs, or other identifiers that allow users to access digital content.
33.	No of pages	Numeric	Eg- No of pages of one file or volume.
34.	Bundle No.	Numeric	---

Annexure -A

Descriptive list of the Meta data fields:-

1.	**Bundle Barcode-** B Barcode represents the bundle barcode number that is affixed to the bundle or volume. It is a unique 14-digit number that contains information pertaining to a specific bundle and all the files linked to that bundle.
2.	**File Barcode-** F barcode is a 14-digit unique barcode number that is affixed to files and contains information about the specific file.
3.	**Location-** Location refers to the physical location or storage site where records or documents are housed or stored within an organization. This information is important for tracking the physical whereabouts of records, managing access to them, and facilitating retrieval when needed.
4.	**Location Code- It's a 14 Digit unique code representing as follows:-** **1.** Room Master Numbers (Three Digit): This is a three-digit number used to identify a specific room within the Stack area. For example, "003" represent Room 003. 2. Compactor or No compactor (One Digit): This single-digit number likely represents a specific compactor within the room. If compactors are there 1 will be entered and in case no compactors 0 will be marked. 3. Pillar Number or Compactor Number (Three Digit): This three-digit number refers to the pillar number or the compactor number as a specific identifier for the room. 4. Rack Number (Three Digit): This three-digit number indicates the specific rack within the compactor or pillar where the items are stored. 5. Shelf Number (Two Digit): This two-digit number further refines the location to a specific shelf on the rack where the items are stored. 6. Location of Bundle or Volume (Two Digit): This two-digit number will represent a specific location within the shelf where the bundle or volume is stored. Example breakdowns: - Location 1: Room 003, Compactor 0, Pillar/Compactor 005, Rack 001, Shelf 12, Bundle/Volume 01 - Location 2: Room 013, Compactor 1 , Pillar/Compactor 005, Rack 001, Shelf 12, Bundle/Volume 02 003-0-005-001-12-01 013-1- 005-001-12-02

5.	**Identifier – It's a 14 digit unique code provided to the documents while uploading on the Abhilekh patal.**
6.	**Category –** Category typically refers to a classification or grouping that helps organize and describe a dataset, records, or piece of information. Categories are used to group similar items together based on shared characteristics or attributes, making it easier for users to discover and navigate through datasets or records. Including a "Category" field in metadata provides a high-level overview of the content or theme of the dataset, making it easier for users to search, filter, and browse through datasets based on their interests or needs. By assigning relevant categories to datasets, organizations can improve the discoverability and usability of their data resources e.g.- Categorization of records as Public records, Private Archives, Cartography, oriental records, Microfilm, Oral Archives etc.
7.	**Status of digitization – The document is digitized or not digitized.**
8.	**Title-** Title in the metadata fields serves as a key piece of information that helps describe and identify the records. It is used to provide a concise and meaningful name for the document, making it easier for users to locate and understand the content of the record. The document title plays a crucial role in organizing and retrieving information within a collection of records, allowing users to quickly identify the document they are looking for based on its title.
9.	**Keywords –** Keywords are used to capture the main topics or subjects covered by the record. They help improve the searchability and discoverability of records on the web or in digital collections. Keywords in metadata can help users and search engines better understand the content of a resource, making it easier to find and retrieve relevant information.
10.	**Scale-** Scale refers to a property or attribute that describes the level of detail or resolution at which geographic data or maps are represented. For example, in geographic information systems (GIS) or cartography, scale is an important concept that indicates the relationship between the size of a feature on a map or in a dataset and its actual size on the Earth's surface. Scale can be represented in different ways, such as verbal scale

	(e.g., 1:10,000), representative fraction (e.g., 1/10,000), or as a scale bar on a map. In metadata fields related to geographic data, the scale information helps users understand the level of detail captured in the dataset or map. It provides important context for interpreting and using the data effectively, especially when combining different datasets or analyzing spatial relationships. Including scale information in metadata fields associated with geographic data sets or maps is important for proper data management, interoperability, and ensuring that users have the necessary information to make informed decisions about using the data.
11.	**Ministry/ Department/ Residency –** Ministry, Department, and Residency information in metadata fields associated with government datasets or resources helps provide context about the origin, ownership, and relevance of the data. It allows users to identify the responsible government entities and understand the administrative context within which the data was produced or managed.
12.	**Branch –** Branch typically refers to a specific division, unit, or branch of a department, agency, or organization. Government departments or agencies often have multiple branches or divisions that are responsible for specific functions, tasks, or areas of expertise within the larger organization. Each branch may focus on a particular aspect of the department's work, such as policy matters, operations, research, or service delivery. In metadata fields, the "branch" field may be used to specify the particular branch or division of a government department or agency that is associated with a dataset, records, or piece of information. This information helps provide context about the specific unit within the organization that was responsible for creating, managing, or distributing the data. Including branch information in metadata fields can help users identify the specific organizational structure within a government department or agency that is relevant to the data or records they are accessing. It can aid in understanding the organizational structure, responsibilities, and expertise of the branch that produced or is associated with the information.
13.	**From Year - Date(mm-dd-yyyy)- To year Date(mm-dd-yyyy)** 1. From year: This field typically represents the starting year of a specific time period or range. It indicates the beginning point or starting date for the data or information being described. For example, if a dataset covers the period from 1748 to 1765, "From year" would be 1748. 2. To year: This field represents the ending year or final year of a specified time period or range. It indicates the concluding point or end date for the data or information being described. Using the previous example, if a dataset covers the period from 1748 to 1765, "To year" would be 1765.

	Including "From year" and "To year" dates in metadata fields help users understand the temporal scope or duration of the data or resource. It provides context about the timeframe covered by the information and can be useful for researchers, analysts, and other users who need to know the specific time range associated with the dataset. By specifying both the starting and ending years in metadata fields, users can easily determine the period covered by the data and make informed decisions about the relevance and applicability of the information for their needs.
14.	**Geographical References-** Geographical references typically refers to a location, geographic area, or spatial extent associated with a dataset, resource, or piece of information. This metadata field provides information about the geographical context of the data, helping users understand where the data is relevant or applicable. The " Geographical references" field can be used to describe various types of locations or geographic features, such as: 1. Administrative boundaries: This may include countries, states, provinces, cities, municipalities, or other administrative divisions. 2. Physical locations: This may refer to specific geographic coordinates, landmarks, buildings, natural features, or regions. 3. Geographic regions: This can include broader geographic regions, ecosystems, climate zones, or other spatial categories. 4. Spatial extents: This may describe the extent or coverage area of a dataset, such as a study area, survey region, or geographic boundary. Including "Place" information in metadata fields helps users understand the geographic context of the data and where it is relevant or applicable. It can aid in spatial analysis, mapping, and decision-making processes by providing information about the location or area covered by the dataset. By specifying the relevant "Place" in metadata fields, users can easily determine the geographic scope of the data and assess its suitability for their particular needs or research purposes. This information is especially important for geospatial datasets, maps, and other resources where location plays a significant role in the interpretation and use of the data.
15.	**File No. /Reference no./ Sheet No./Folio no.**

	File number or reference number/ sheet no/Folio no. in metadata fields typically refers to a unique identifier or reference number assigned to a specific file, document or maps. This identifier helps to distinguish and track individual files within a system, database, or collection. File numbers are used to uniquely identify files and facilitate organization, retrieval, and management of documents or data. They can be alphanumeric codes, numerical sequences, or a combination of letters, numbers, and symbols. Including a "File number/Reference number sheet no/Folio no " field in metadata allows users to quickly locate and reference specific files within a dataset or document repository. This unique identifier helps in efficient file management, version control, and tracking of documents, especially in environments where multiple files are stored and accessed regularly. By assigning file numbers/reference number to documents and incorporating them into metadata fields, organizations can streamline the document management processes, improve searchability, and ensure accurate identification and tracking of files within the system.
16.	**Part Number** Part Number in metadata fields refers to a classification or categorization of a specific part or component of a file, volume or dataset. This metadata field helps to describe the nature or role of individual parts or components and provides additional context about file/volume.
17.	**Physical condition of the record.** Physical condition of the records in metadata fields refers to the state or quality of the physical medium on which the records are stored. This information helps to assess the preservation status and potential usability of the records over time. When describing the physical condition of records, metadata may include details such as: 1. Condition: Describing if the records are in good, fair, or poor condition. This could include factors like wear and tear, damage, deterioration, or fragility. 2. Format: Indicating the physical format of the records, such as paper documents, photographs, microfilm, magnetic tapes, optical discs, etc. 3. Extent of damage: Specifying any specific damages present, such as tears, stains, mold, fading, scratches, or other forms of physical degradation. 4. Stability: Noting the stability of the physical medium and any risks of further deterioration or loss over time.

	By including information about the physical condition of the records in metadata fields, organizations and archives can make informed decisions about the appropriate preservation strategies, conservation efforts, and storage conditions needed to safeguard the records for future use. This information also helps users understand the quality and reliability of the records and may influence how they handle, access the materials for long-term preservation.
18.	**Document last date of repair-** Including the last date of repair in the meta fields of archival records helps in tracking maintaining history, ensuring the integrity of the records, identifying the need for conservation or restoration, facilitating compliance with archival standards and adding in decision making related to the preservation of the records.
19.	**Series** In metadata fields, "Series" refers to a grouping or collection of related items, documents, or records that are organized together based on a common theme, topic, or characteristic. A series helps to structure and organize information in a systematic way, making it easier for users to identify and access related content. In archival and library contexts, a series often represents a set of records or documents that are closely related in terms of content, function, or source. Each item within a series is part of a larger group that shares common attributes and is typically arranged in a specific order or sequence. Eg- QMG(Quarter Master General Series), WWI- First World war diaries etc.
20.	**Source organization-** It refers to the organization or institution responsible for keeping or contributing to the records described by the metadata. It is a metadata element that helps identify the entity or entities associated with the creation, maintenance, or dissemination of the records. The "organization" field can be populated with the name of the organization or institution, such as a Department, company, university, government agency, or any other entity responsible for the records. It helps users understand the source and authority behind the information provided for the records and can be valuable for purposes such as attribution, citation, and management. Eg- National Archives of India.
21.	**Record Contains** The term "record contains" typically refers to a description of the contents or the specific elements within the records being described. This metadata element is used to provide information about the actual content found within the records itself eg- Maps, photographs, treaties, seals, coins, currency, illustrations etc.
22.	**Language** Language element is used to indicate the language of the intellectual content of the records being described. This field helps users understand in

	which language the content of the resource is primarily written or presented. By including the "language" field in metadata records, creators and users of resources can easily identify the language in which the records is written, spoken, or presented. This information is particularly useful for multilingual users who may need to search for records in specific languages or for systems that support language-based filtering and retrieval.
23.	**Microfilm Roll Number** In the context of archival or library metadata, the "Microfilm roll number" is a specific piece of information used to identify a microfilm reel within a collection. Microfilm rolls are often used to store copies of documents, newspapers, photographs, and other types of materials in a compact and long-lasting format. When cataloging or describing materials that are stored on microfilm, archivists and librarians assign a unique identifier to each microfilm roll to facilitate organization, retrieval, and access to the content. The "Microfilm roll number" serves as a reference point to locate the specific reel containing the desired information within a larger collection of microfilmed materials. Including the microfilm roll number in metadata records helps users and researchers quickly identify and access the relevant microfilm content they are looking for.
24.	**Relation** Relation field is used to establish relationships between the resource being described and other material . This element helps provide context and connections between different resources, allowing users to navigate related materials and understand how they are interconnected. Relation field can be used to indicate various types of relationships, such as: 1. Is part of: Indicates that the resource is part of a larger collection or series. 2. Has part: Indicates that the resource contains smaller parts or components. 3. Is version of: Indicates that the resource is a different version or edition of another resource. 4. Has version: Indicates that the resource has different versions or editions. 5. Is referenced by: Indicates that the resource is referenced by another resource. 6. References: Indicates that the resource references another resource. By including the "Relation" element in metadata records, creators and users of resources can explore connections between related materials, navigate through collections, and discover additional resources that may be

	of interest. This field helps enhance the discoverability and usability of digital resources by providing valuable context and links to related content.
25.	**Copyrights** The "Copyrights" field in metadata is used to provide information about the copyright status and restrictions associated with the resource. This may include details about the copyright holder, copyright date, usage rights, licensing information, and any restrictions on how the resource can be used, copied, or distributed
26.	**Call Number** Call Number is a unique identifier assigned to a resource in a library or archive to facilitate organization and retrieval. It helps users locate the physical or digital item within the library's collection
27.	**Publisher** The "Publisher" field identifies the entity or organization responsible for publishing, issuing, or distributing the resource. This information helps users understand the source of the resource and can be important for citation purposes
28.	**Subject** The "Subject" field describes the topic or topics covered by the resource. It helps users understand the content of the resource and facilitates the discovery of materials related to specific subjects.
29.	**Creator** This field identifies the person, organization, or entity responsible for creating the resource/book/material. This element helps attribute authorship or creation and provides information about the origin of the resource/book/material.
30.	**Accession Number** It is a unique identifier assigned to a resource when it is acquired by a library or archive. It helps track and manage the resource within the collection.
31.	**Year of publication/ (YYYY-MM-DD)** The "Year of Publication" element specifies the year in which the resource was published or made available to the public. The optional inclusion of the exact publication date (YYYY-MM-DD) provides more precise temporal information.
32.	**ERA** The "ERA" (Electronic Resource Access) element may refer to information related to accessing electronic resources, such as URLs, DOIs, or other identifiers that allow users to access digital content.

Annexure-B

Descriptive List of Standardized Location of Records.

This is a descriptive listing of the master room numbers which will be incorporated with the location code format for all the repositories to streamline the process of sharing location details of all the available documents in the process of digitization.

Specifically, a **"Horizontal Drop box Menu"** can be implemented of the following format featuring -

Room Master Numbers	**Compactor**	**Pillar Number or Compactor Number**	**Rack Number**	**Shelf Number**	**Location of Bundle or Volume**
XXX	**X**	**XXX**	**XXX**	**XX**	**XX**
001-999	0 or 1 If compactor is available, it will be assigned the value 1 and if it is not available 0 will be assigned.	001-999	001-999	01-99	01-99

UNIQUE LOCATION CODE - 14 Digits (Numerical)
Eg:- if the compactor system is available, the location code will be **00310050011206**
If it is not available, the location code will be**- 00300050011206**

Room Master Numbers [(XXX) 3 Digit Code]

These are assigned to rooms where records are stored. Room Number has been allocated to existing repositories, with future expansion in mind to accommodate additional repositories including all ROs / RCs.

Room Master Numbers may be assigned as follows-
001-200 - National Archives of India, Delhi Headquarters
201- 300 - Lahore Shed
301- 400 - Bhopal, RO
401-500 - Jaipur
501-600 - Bhubaneswar
601- 700 - Puducherry

Compactor [(X) 1 Digit Code]

These are assigned to rooms where records are stored to existing Compactors or Pillars,

1- If the Documents are stored in Compactor,

0- If the Documents are not stored in compactors.

Pillar Number / Compactor Number [(XXX) 3 Digit Code]

To accurately identify the location within the **NAI building**, please input the corresponding **three-digit code** for either the pillar number or compactor number, based on your selection in the previous menu.

If you selected 1 in the previous menu, please enter the compactor number within the range of 001-060.

If you selected 0 in the previous menu, please enter the pillar number within the range of 001-128.

This information will facilitate precise navigation within the building premises.

Rack Number [(XXX) 3 Digit Code]

This three-digit unique code serves as a direct access key to the designated rack where all identified and necessary documents are stored. It facilitates precise identification of the appropriate rack for accessing specific sections of documents during search processes.

Shelf Number [(XX) 2 Digit Code]

Following the selection of the correct rack number, the two-digit unique code assigned as the shelf number enhances clarity and facilitates ease of access to locate the exact document shelf containing the required documents.

This systematic approach optimizes document retrieval processes by guiding directly to the designated shelf, thereby minimizing search time and improving efficiency in document access.

Location of Bundle / Volume [(XX) 2 Digit Code]

This represents the final step in identifying and accessing the required documents. After inputting all the correct details, rack number and shelf number, precise access to all documents is facilitated by entering the truth-rich code for the location of bundles and volumes. This step provides exact location of the volumes and bundles available within the repository , ensuring efficient and accurate retrieval of the desired documents.

This above mentioned standardized location code format aims to enhance efficiency and accuracy in locating documents, it will also streamline the document management system in NAI and improve the overall experience for the staff members.

Example of bifurcating Master Room Number for National Archives of India, Delhi Headquarters -

Repository	Master Room Number	Location in NAI	Numbers of Compactors /Pillars	Remarks
R4	001		128 pillars	
R4	002	Gangways	128 pillars	
R2	003		128 Pillars	
R2	004	Gangways	128 Pillars	
R3	005		128 Pillars	
R3	006	Gangways	128 Pillars	
R4	007	Annexe Building Second floor		
PA	008	Annexe building , third floor (2 Rooms)		
PA	009			
Cartography	010	Annexe building , third floor (2 Rooms)		

Comments on Draft Guidelines for Digitization for Records received from five State Archives/Institutes.

#	Institution	Comment/Suggestion		Remarks by NAI
1	Delhi Archives.	1.	SOP for digitization of archival records seems to be high standard in Indian condition, and most of the State/UTs are not well aware the basics of digitization. Delhi Archives has suggested to formulate two types of standard (1) one for beginners Archives and (2) second for advance Archives.	This SOP (Standard Operating Procedure) provides a set of best practices and guidelines for production of digital records and its management. It is applicable for those digital records that need to be retained for long durations and need to be preserved permanently. The core concepts of this SOP are based on the practical experience gained by NAI through their past digitization projects and globally available best practices and guidelines for digitization. The objective of this SOP is to formulate a uniform practice in digitization so that all the data created by other record keeping agencies should be available on single platform for the users in future. The record keeping agencies may adopt this SOP as per the availability of infrastructure. The accuracy
		2.	The technical standards for image capturing are quite same which have already been used by the Delhi Archives under its Project of Digitization and Microfilming of Archival Records.	
		3	OCR of Handwritten Recognition need to be reviewed considering the provision of 50% accuracy prescribed in the draft SOP, which need to be reviewed and enhanced at least up to 90 to 95%.	
		4	OCR of archival documents need to be reviewed at this stage, where many of the archives is not even started the digitization of archival records and not even having any IT professional.	

2	Kerala State Archives	1	Regarding the OCR recognition and for the development of transcription software, the main hurdle is the familiarization of old scripts to machine. Since it is a huge task to learn the machine regarding the all possible variants, need much time and effort for collecting such data. So it's better to develop a semi automatic software which serve as a common platform for the transcription many old regional languages. So a better solution regarding this is also expected. The suggestions available from the NAI will be beneficial to Kerala State Archives in many means. This may helpful while executing various activities undertaken by and can switch to a better system by incorporating all these suggestions.		Handwritten Recognition is still in the stage of developing so accuracy up to 50% is fine at initial stage of digitization as the concept of AI (Artificial Intelligence) has also has come which is working on creating the data set of various types of handwriting in Multilanguage will enhance the accuracy of OCR up to 95% in future.
3	Anandashram Sansthа, Pune		Anandashram Sanstha, Pune is a recognized Manuscript Resource Centre (MRC) from GOI. They have digitized 600000 pages of mss and 15000 additional pages of mss are yet to be digitized. NAI efforts for standardization are very commendable but in order to be successful in implementing, it will need 2-3 years of separate efforts on their part. They are seeking Substantial additional funding from NAI so that they can submit a draft project proposal by August end if NAI agree to provide fund.		N/A
4	Bihar State Archives		NAI SOP Points	Comments of Bihar State Archives.	The Bleed through ink has to be removed before doing HWR. The sharpness of ink should be enhanced during image processing. The HWR technology is using the dataset prepared
		1	In case of handwritten documents, handwriting recognition	Need more clarification in case of : (1) Significant	

			should be of carried out. At least 50% accuracy has to be ensured in HWR and multiple language support should be supported. [Point-4(5)]	fade of the ink and bleeding of text from back side of the page. (2) Variety of writing styles abbreviations, symbols.	from the different partners, curves and stroke of writing available in different types of handwriting and symbol to give result. The NAI is in process of implementing the AI based auto tagging which is in initial stage the final tested output will be shared in due course of time.
		2	AI based auto tagging should be carried out to generate keywords for all scanned documents. These keywords will facilitate more meaningful search operations than tagging. merely looking at the subject lines of documents. [Point-4(7)]	Please provide some guidelines for use of AI based applications or process which supports the AI based auto tagging.	
			The OCR/ HWR text extracted from each page, auto tags generated and metadata should form part of the PDF/A file. It would be	Please share a common format of quality check certificate.	

			desirable to include the quality check Certificate also within the PDF/A file. [Point-5(3)]		
5	Karnataka State Archives		The Karnataka State Archives had gone through the draft Guidelines for Digitization of Records and it is observed that with regard to bound scanning of bound volumes it is better to unbound the volumes and stitch the volumes after scanning. Even though cradle type of scanners should be used for scanning, there is a chance of block mark in the middle of scanned document which can not be visible for the reader.		The bound volume should be scanned by the v type overhead scanner which do not required the un bounding of big volume.