Recommendations for Digitization

1. Items to be considered

From the first book and periodical - all the items that were published, including official publications, pamphlets, posters, Palm leaf manuscripts, paper manuscripts, diaries; Potteries and other antiquities from the archaeological excavations and explorations that includes, potteries, inscribed potteries, terracotta objects, ornaments, beads, metal objects, shell objects, bone objects, structural remains, coins, all bronze images **with authentication,** Paintings, Murals, rock cut architecture and the sculptures, copperplates containing inscriptions should be digitized.

In addition to this, annual reports on Epigraphy right from 1880, South Indian Inscriptions ancient volumes, Ancient India volumes, Memoirs of Archaeology, Indian Antiquary volumes – these are some of the ancient reports on archaeological findings. In case of identification of any new artifacts it shall be notified to the director of the TVA.

2. Cataloguing

1. Identify collections and catalogue the unique items
2. Printed catalogues may be digitized
3. A Union catalogue is required for Tamil collections
4. MARC format shall be followed as a standard to catalogue items
5. Items shall be catalogued in the language of the imprint
6. Proper keywords, taxonomy and ontology have to be imposed in catalogues
7. Transliteration in Roman script may be implemented to enable scholars in other countries
8. ISO 15919 standards may be used to transliterate
9. If a catalogue is already developed by an Institution, an MoU may be drawn to provide links or the data to be viewed through TVA platform
10. Cataloguing standard for paintings, developed by French Institute of Pondicherry and American Institute of Indian Studies, Chennai may be considered for cataloguing.

A similar approach may be followed for palm leaf manuscripts, paper manuscripts, epigraphically works, rock art and other material listed in section 1. Wherever required, a unique approach may be developed without reinventing the wheel. If there is no protocol for an object/ artifact, then new standards may be developed.

3.1 Selection of books for preservation

1. Books published from 1556 onwards up until 1957 can be taken for digitization immediately
2. Books that have been nationalized could be taken up for digitization
3. Items published after 1957 and for those items that does not have a copyright owner shall be identified and digitized
4. Orphaned works could be digitized with proper disclaimers *(However this may require an opinion from legal experts)*
5. Copies of rare items from other countries may be brought in a similar approach can be taken for periodicals
7. Include Texts books, political pamphlets without any bias
8. Contemporary works such as novels, short stories, manuscript of authors, letters or authors, diaries may also be brought in with permission from copyright owners
9. In the case of periodicals and journals, they need to be catalogued and indexed

3.2. Selection of Manuscripts

All manuscripts have to be catalogued and digitized. An electronic union catalogue may be planned.

3.3. Selection of other materials

All items need to be selected for the purpose discussed here. However, prioritization may be decided based on funds and expert advice.

4. Precautions

1. Inventory of digitized material has to be taken
2. Duplication of work to be avoided
3. A committee could be formed to prioritize the material for digitization. However all items need to be digitized in a phased manner
4. If there is a digital copy already available, they may be either purchased or brought in through some understanding.

5. Digitization Standards

ISO 14721:2012 Open Archival Information Systems may be taken as a reference model. Standards laid by Library of Congress, Endangered Archives Programme, Project Athena, Research Library Group may be followed for digitization. There is no need to reinvent the

wheel.

360 degree scanning may be considered for artifacts.

6. Using OCR

OCR for books, periodicals, epigraphicals works and other texts could be taken up as value added service for searching data. However, this should produce at least 99.5% error free data. If this is not possible then more funding may be provided for researching and developing such a product. In the case of epigraphical works, it may be worth typing the data instead of running through an OCR.

7. Access

1. An online access may be provided through the TVA or associated agencies or collaborating partners
2. Access may be provided as graphic images and in epub formats
3. For those items funded by TVA, they may be placed on public domain for open access
4. For those items not funded by TVA but partnering institutions willing to provide data may also be placed on public domain following due diligence.
5. Fair usage policy with disclaimer may be used for access of the digital content
6. Watermarks can be inserted for every page and user registration may be used for access

8. Other recommendations

1. An implementing agency (IA) may be set up which could be a different agency or the Tamilnadu Virtual Academy itself as designated by the Government
2. A copy of archival quality digital data has to be stored in the Government data centre and partner institution
3. Owners of the physical and digital material may be fairly compensated and duly acknowledged
4. Standards, protocols, processes and methodologies have to be made available in Tamil
5. Working groups may be formed for each area to work under the auspicious of TVA

9. Functions of Implementing Agency

1. Survey of already digitized contents
2. Collecting digital content from various Government departments, private libraries and individuals
3. Uploading the digital content along with metadata
4. Maintaining and updating the digital content in a server
5. Migrating the digital content to new formats
6. Providing security to the digital assets

Some of the above mentioned areas may be designed into projects and stake holders may be invited for collaborations.

These recommendations may be further elaborated into processes for each of the activities. Book banks may be created in repository libraries in addition to physical copies. This may require legal opinion.

Perishability of digital records; testing OCRs for performance

1. OCR Testing

The OCRs need to be objectively evaluated for their performance: A set of test documents (8-bit gray images scanned at 300 dpi) must be created with at least a few pages from each decade, starting from 1800 to 2000. This will obviously cover noisy pages and old letters of Tamizh, like யானைக்கொம்பு, etc. Ground truth needs to be created for all these pages. All available OCRs should be tested on these pages, to objectively determine their capability.

2. OCR - Standardization of requirements:

It is recommended that we come out with a document, specifying what the requirements of a good OCR is, in terms of input and output requirements.

3. Perishability & Expensive nature of digital storage:

Digital records are HIGHLY PERISHABLE, unlike books and palm leaf manuscripts. They need constant upgradation of hardware (storage media), software (format of storage such as .tiff, etc.), mirroring, etc. and hence is quite an expensive thing to maintain for long. So,there must be a policy from the Government to assure long term support for the infrastructure, maintenance and upgradation. It also needs qualified IT staff to manage.

4. Optimal resolution of images:

Resolution of images higher than needed will unnecessarily increase storage and also access time for users.

-AGR