Digitization Technical Requirements

Technical specifications aligned to the digitization requirements should be documented and implemented. The quality of equipment and software used in digitizing significantly affects the capability to support appropriate technical standards. The hardware chosen for a digitization project should be maintained according to a regular maintenance program. Scanner equipment should meet the specifications required by PASI for Digital Student Record.

All source records should be scanned at a minimum of **300 DPI**, and must be output to **PDF**.

Compression - reduces the size of the digitized record to enable efficient storage and easier transmission. Digitized records may be lossless or lossy:

- **Lossless compression** reduces the size of the digitized record, to approximately half its original size, with no loss of quality and is preferred for high quality reproduction purposes.
- **Lossy compression** is suitable for digitized records that don't require a high quality reproduction such as photographs where minor loss of accuracy is acceptable to achieve a substantial reduction in size.

The technical specifications of acceptable formats as per GOA Digitization Technical Requirements Standard are:

Class	**Tonality**	**Bits**	**Compression**
Black and white documents or Microform	Bitonal	1	Lossless
Colour documents	Bitonal or Color	24	Lossy or Lossless

Greyscale works great for handwritten records because details display better than a bitonal scan.

Quality Control of Source Documents

The purpose of Quality Control is to ensure that the digitized record mirrors the physical record. Errors can occur during digitization due to scanner mis-feeds or poor quality physical records. In order for the documents to proceed through an appropriate and auditable imaging procedure, the items should undergo Quality Control checks in the following areas:

- Document selection
- Document preparation
- Quality of the scanned/re-scanned product

Quality checking should include ensuring the readability of the output images and that the number of physical items scanned is equal to the number of images produced. Any enhancements of the image (e.g. de-speckling, de-skewing) should be used with caution as they may be deemed as altering the original document by the courts.
Quality Control techniques may include:

- Count the number of pages of the digitized records and ensure they match the number of pages of the Physical Records. Be aware of blank pages in the digitized record that can count as a page
- If multiple documents are digitized in a single batch, capability to separate individual digitized documents should be assured
- Ensure images are in the correct order
- Ensure page alignment is correct e.g. physical record orientation (portrait/landscape), rotation, image skew, cropping etc.
- Check for completeness and accuracy of detail e.g. readability, text clarity, sufficient capture of punctuation marks, etc.
- Check for scanner generated speckle e.g. speckle not on the original document
- Check for density of solid black areas. An example of where this is problematic is if a physical record has highlighted text, the digitization could black out the text
- Remove blank pages

Metadata/Indexing Requirements

Metadata should be captured and managed to prove that records are complete, accurate and trustworthy. Metadata/indexing should be retained for at least as long as the records to which they relate are retained. Wherever possible metadata should be inherited from system. The digitization process includes four phases where Indexing should be applied. These phases are:

- Image capture (scanning)
- Image re-capture (re-scanning)
- Quality Assurance
- Data Transfer

There are two types of Indexing information:
Biographical information deals with the lifecycle of the image file, and relates to the context of the image and file properties that should be captured, logged, and certified during the digitization process.
Bibliographical information relates to the content and context of the record. This information should be captured and then associated with the image, preferably by automated means or by manual data entry during the digitization process.

PASI Metadata Elements

The following are the metadata requirements for loading of documents into PASI as per section 2.3.1 in PASI Readiness Overview:

- Alberta Student Number – a number used to uniquely identify each student.
- Document Type – identifies the type of document that is being added to the student (must align with a predefined list of acceptable document types within PASI)
- Title – the commonly known as name for the document from the user's perspective
- Relevance – a true or false value that indicates if the document is relevant within the complete student record and should be reviewed as part of the initial student record review
- Document Language – indicates the language of the content of the document (English, French, other)
- Document Date - represents the date that the document was generated on or attributed to
- Document Expiry Date – the date the document expires.
- Linked to School Year – reflects that school year that the document pertains to.
- Quality Assurance Already Performed – a true or false value that indicates if quality assurance has already been completed on the document
- Original File Name – the original file name as it was loaded to PASI
- Linked to Organization – a K to 12 organization that the document belongs to and is linked to in PASI

- Text Searchable – a true or false value that indicates if text character recognition and searching is available for the document
- Digitized – a true or false value this indicates if the electronic document was created from a hard copy document
- Exempt from QA – a true or false value that indicates the document is considered exempt from quality assurance

Re-Capture Image

The re-capture is required if the images and associated Indexing fails the quality control. When errors are found in the initial digitized record, it is mandatory that the physical record be re-digitized. Upon re-digitization of the physical record the digitized record will have to go through the quality control step again and the process repeated until no errors are detected.

Optical Character Recognition (OCR)

Optical Character Recognition is the process of converting digitized records into machine-encoded/computer-readable text. This process allow digitized records to be searched using keywords.