TDR Digitization Standards and Procedures

This document describes the standards and procedures the CRKN uses for digitizing, packaging, performing quality control, and ingesting material into the Canadiana trustworthy digital repository (TDR). It is meant to provide depositors and users of repository content with an understanding of the requirements for inclusion of material in the TDR, the processes by which material is created and processed, and the potential limitations of digital representations within the repository.

This document may be distributed publicly. The current version of this document is located at: https://docs.google.com/a/c7a.ca/document/d/10Q03FCFJpzWjLg_HM2fFaAkCdJBaKWyvqC_L1qthbRA/edit?usp=sharing

Source and Format of Material

CRKN accepts material for ingest into the TDR that CRKN itself has digitized according to its own production and quality control standards. It also accepts material that has been digitized and prepared by a third party depositor. Depositor-generated material may either be in the form of a complete submission information package (SIP) or it may be in the form of incomplete, unpackaged or partially-packaged material that must be further processed by CRKN before it can be ingested.

The following outlines the procedures for CRKN-produced digital content and notes differences where depositors supply all or part of the digital material for ingest.

Document Preparation

Original Documents

The general procedure for preparing paper documents (e.g. books, pamphlets, maps, photographic prints) is as follows:

1. Check the received documents against the packing list to ensure each document is present in the shipment. Also, if necessary, a list is created for all of the documents that have been placed in a box. Number each box and record the box numbers on the document list. Make a copy of the list for tracking purposes.
2. Unless directed otherwise by the client, remove all clips, brads, staples, and other fasteners from folders, books and from multi-page documents. Use an appropriate tool to prevent tearing or otherwise damaging the documents.
3. Carefully unfold and flatten documents to eliminate creases and wrinkles.
4. Check every document to make certain that all documents are available or accounted for and that they are in their proper sequence.

5. Check the numerical sequence and the order of all pages and note all omissions; note missing parts of pages, stains, tears, or obliterations that affect the text of any document; and note all other irregularities that affect the legibility of the material. On the list of documents write the number of pages for each document.
6. Remove sticky notes and flags, unless they contain important information. If sticky notes and flags must be retained because they contain important information, leave them in the document to be removed as the document is scanned and then replaced on the document; or tape onto a separate sheet of paper so that they do not obscure the main document.
7. Identify documents to be enhanced by using special scanning techniques.

Microform

The general procedure for preparing microform documents (microfilm and microfiche) is as follows:

1. Check the received microforms against the packing list to ensure each microform is present in the shipment. Also, if necessary, a list is created for all of the reels/microfiche that has been placed in a box. Number each box and record the box numbers on the list of microforms. Make a copy of the list for tracking purposes.
2. Sort the microforms by reduction ratio before scanning. If the reduction ratio changes within the reel a new scan needs to be done for those images using the new reduction ratio.
3. Check the numerical sequence and the order of all pages and note all omissions; note missing parts of pages, stains, tears, or obliterations that affect the text of any microform; and note all other irregularities that affect the legibility of the material.

Scanning

Materials are scanned on hardware appropriate to their format and condition. Specific scanning procedures will vary according to the particular hardware. For all types of scanning, the general process used to create digital images from analog originals is as follows:

1. Each day, before using any scanner for production work, perform a test of the scanner by scanning a sample of content from the next production batch to be processed on that hardware, using the same settings that will be used to process that batch. Compare the test output to quality reference images. Scanners that produce images that do not meet quality standards must not be used for production work until they have been properly adjusted, calibrated, repaired or otherwise corrected.
2. Scan documents in the working batch using the appropriate settings for that batch. As objects are scanned, return them to the original box or other packing container.

File Formats and Quality Standards

Technical Standards

The standard default file format for digitized still images is a JPEG image with high-quality compression (quality level of 80%) with an optical resolution of 300 dpi. Images scanned from print sources are saved as 24-bit colour. Images scanned from microform sources are saved as 8-bit greyscale.

Other resolutions, colour depths, compression schemes and file formats may be used provided that the resulting file format is supported by the Canadiana submission information package (SIP) standard and that the resulting image still meets general quality guidelines.

Image Quality

In order to pass quality control, images must meet the following general criteria:

1. Images must be complete (no missing frames or pages).
2. Images must be in the correct order.
3. No duplicate images must be present.
4. Images must be properly oriented. Images can either be in their original orientation or can be rotated to their natural reading orientation but may not be upside down, mirrored or in any other orientation.
5. The content of the images (text, illustrations, etc.) must be clearly legible or viewable. Scans of damaged, stained or poor-quality originals are acceptable, but the accompanying metadata should include a note indicating the presence of such damage or other source image problems.
6. No moiré patterns, visible pixelation, lack of focus, or other artefacts which indicate problems or defects in the scanning process appear.
7. If images are cropped such that the edges cannot be seen, that no significant content is missing from the edges, such as page numbers or marginal notes.
8. The dynamic range of the image (difference in brightness between the darkest and lightest components) is great enough to facilitate further manipulation for purposes such as OCR and creating legible derivatives.

Cataloguing and Metadata

Each document requires an associated metadata record to identify and describe it. A variety of metadata formats and standards are acceptable, depending on the nature of the source material, provided that it meets the technical and quality requirements for inclusion in a Canadiana SIP. CRKN creates metadata records using a variety of current descriptive cataloguing standards, which include Dublin Core, Resource Description and Access (RDA),

and MARC21. Metadata is created in-house by qualified cataloguers, or by third party depositors in coordination with CRKN staff.

Quality Control and Post Processing

CRKN performs a quality review of a sampling of digitized images in which an operator checks random images and compares it against the image quality standards described above. For print materials, the digitized images are proofed against the original item as a reference. Microform materials go through a post-processing step where a technician visually inspects the scan of the entire microform and ensures that all exposures on the film or fiche are properly detected by the software, making manual adjustments as needed before performing the final step of outputting individual images for each frame in the original.

Quality problems detected at this stage are resolved either by manual correction (for example, rotation of images, deletion of extraneous images, or rotating of skewed images) or by re-scanning missing or poor-quality images.

CRKN requires that depositors who submit already digitized images perform quality control processes and adhere to standards comparable to those followed by CRKN. In addition, CRKN performs the following quality control analysis on digital images received:

1. A random sample of images are inspected and evaluated according to CRKN's image quality standards.
2. A profiling tool is run against each batch of images to detect anomalies such as unexpected file formats, unusual numbers or sizes of images, and other elements that might indicate problems. Based on this, additional images may be flagged for inspection and evaluation.
3. Metadata records received from a depositor are reviewed by comparing a random sample of supplied records against the supplied source documents (originals or digitized images). This analysis consists of comparing the content of the record with the chief source of information to verify that the description provided in the record reasonably approximates the corresponding material.

A minimum of 90% of the documents inspected must meet CRKN's image quality standards. Batches that do not meet this pass rate will be rejected as having an unacceptably high failure rate. All detected failures must be corrected, either by CRKN or by the depositor, before further processing of the batch will take place.

OCR

CRKN typically applies OCR processing to digital images with printed text. OCR processing is not required and is selectively applied based on the nature of the content, the project to which it belongs, and agreements made with any clients or third parties. All OCR is "dirty OCR" meaning it is not manually corrected or post-processed and does not have to meet any particular quality

requirement. While CRKN monitors the overall confidence results of the OCR process for the purposes of ongoing evaluation and fine-tuning of OCR software and methods, it does not keep records or metrics for individual documents or pages.

OCR may include word or phrase positioning (coordinates) or it may consist of plain text, as described in the Canadiana SIP specification. CRKN retains word positioning where it is practical to do so.

Depositor-supplied OCR can be used as long as it meets the general specification for inclusion in a Canadiana SIP. CRKN does not perform any quality control on OCR supplied by depositors.

Packaging and Ingest

CRKN packages digitized content and associated metadata into a standard SIP format. This is done after all digitization and quality control work is completed. Depositors providing their own digitized material may supply their own packaged SIPs or may supply digital components for CRKN to further process and package as needed.

Before it can be ingested into the repository, each SIP undergoes an automated validation process which verifies the following:

- the SIP is a valid BagIt archive
- the checksums in the manifest file match those of the files in the payload directory and that there are no missing or extra files
- the METS record validates against the METS XML schema
- the sub-records (dmdSec elements) are valid and of one of the allowable types
- there is a matching file in the files subdirectory for each file listed in the fileSec of the METS record
- the MIME type of each image file matches its suffix and the file itself is a valid example of that file type (determined using ImageMagick's identification tools)
- the supplied identifier (OBJID) is allowed

A SIP which passes all of these checks will be ingested into the repository. Those that do not must be corrected before attempting to ingest again.

Each SIP is identified by a locally-unique identifier supplied by the depositor. This identifier is used in conjunction with the depositor's CRKN-assigned code to uniquely identify any AIP in the repository. If a SIP is ingested with an identifier that matches an existing AIP's, the SIP will replace the existing SIP as the current representation of that object and the previous SIP is archived within the same AIP as a prior revision. If the identifier does not correspond to an existing AIP, a new AIP is created and the SIP is ingested into it.

References

The definition of a Canadiana SIP is located at:
http://www.canadiana.ca/schema/2012/txt/sip.txt.

The definition of a Canadiana AIP is located at:

http://www.canadiana.ca/schema/2012/txt/aip.txt.

The Canadiana CSIP METS profile (which defines valid file formats for the SIP) is located at:

http://www.canadiana.ca/schema/2012/mets/csip.xml.

Revision History

Version	**Comments**
2015.04.09	Added a note to the packaging and ingest section to specify the use of ImageMagick as the tool used to validate file formats.
2020.03.20	Updated document to reflect CRKN's stewardship of the Canadiana collection
2021.07.30	Updated the following sections: document preparation, scanning, file formats and quality standards