CHM Collection Management and Access Plan

1. Purpose

- The CHM seeks to be the premier international physical and virtual destination for reliable information about the development of computers and the computer industry worldwide. As such, the CHM strives to be the most complete and accessible digital repository of artifacts, documents, media and stories for all aspects of computing.
- In order to achieve these goals the CHM seeks to catalog and digitize the CHM Collection to create the CHM Cyber Collection making the CHM Collection more accessible within the organization and to its target audiences.

2. Audience

- The CHM audience as a whole includes multiple segments, such as researchers, students, teachers, media, computer professionals, vintage computer hobbyists, and interested members of the general public. The CHM Cyber Collection's primary audience will be researchers, students, the media, computer professionals and vintage computer hobbyists.

3. Access vs Preservation

- The purpose of the Cyber Collection is to provide access to the CHM Collection and not to create highly accurate reproductions of the physical form or to replace the physical items as the archival. A secondary benefit of the digitization of parts of the CHM Collection will be to minimize the wear and tear of the physical items and thus assist in their preservation.

4. Collection Database

- The CHM Collections Database will serve all five CHM Collections: Moving Image, Still Image, Artifacts, Text and Software using Dublin Core Metadata fields with qualified CHM metadata fields as required.
- Any item or resource (physical or born digital) that is considered part of the CHM Collection will be given a unique number (Accession Number) and an individual record in the Collections DatabaseThe Collections Database will also link to a variety of "surrogate" images that are related to items entered into the database, such as photos of artifacts; scans of text documents or photographs in the collection. These surrogate images are not considered in the formal sense part of the CHM Collection, as such they will not be given a unique Accession Number nor an individual record in the Collections Database, however they will accessible and forever linked to a specific item in the collection via the database using a pointer in the database record and a related file name.

5. Data Structure & Search Strategies

- The CHM best practice is to capture descriptive information or metadata necessary to uniquely identify each item in the CHM Collection and make it retrievable.
- The CHM has adopted a metadata strategy that will capture descriptive information about each item in both a database (e.g. structured entry record metadata) and in descriptive file folder and file names (e.g. apple_macintosh.jpg).

This data structure strategy will facilitate multiple searching techniques, maximizes access to CHM Collection data. Users will be able to access information through database queries as well as through browsing or text searching file folders and files.

6. Dublin Core Metadata

- The CHM has adopted the Dublin Core Metadata Element Set for its Collection Database descriptive metadata in order to provide a set of common metadata and controlled vocabulary across all CHM collections. The use of metadata in a structured database will allow users to search the CHM Collection most effectively, allowing advanced searching by topic and multiple pointers to related information.
- Dublin Core Metadata was originally developed by DCMI in 1985 to create interoperable metadata standards for use in a variety of settings, including museums, libraries, government institutions, and research institutions, etc. The Dublin Core metadata element set is a national and international standard for cross-domain information resource description.
- The use of an internationally accepted set of Metadata elements will also facilitate collaborative sharing among members of the research community.
- Dublin Core Metadata is flexible, allowing CHM to use only those fields that are relevant as well as providing the ability for CHM to add custom fields as needed, such as the use of non-dublin core qualifiers within each metadata element, e.g. Date_Manufactured; Date_Introduced; Date_Announced, see Appendix I for details specific to CHM collection.
- The Dublin Core Metadata elements include:
    - ID → An unambiguous reference to the resource
    - Title → A name given to the resource
    - Creator → An entity responsible for making the content of the resource
    - Description → An account of the content of the resource
    - Contributor
    - Date
    - Format
    - Language
    - Location
    - Publisher
    - Relation
    - Rights
    - Subject
    - Series
    - Type

7. Database Software

- The CHM has chosen DB/Textworks by Inmagic as its content management software for its Collections Database in the near term.
- DB/Textworks has been chosen for a variety of reasons, including its effective content management tools, simple web publishing tools and minimal cost.

- Although CHM has chosen DB/Textworks by Inmagic for its current needs, the CHM plans to consider future upgrades to more sophisticated relational database software, e.g. SQL as the need requires. The current use of DB/Textworks will not hinder this future upgrade.

8. Digitization Standards

- Since the purpose of the CHM Cyber Collection is to provide access to the CHM Collection, the digitization standards will focus on legibility as opposed to preservation or highly accurate reproduction of physical items.
- In developing the digitization standards for the CHM Collection, the CHM has reviewed the approaches adopted by the following organizations: Library of Congress, National Archives, Museums and the Online Archive of California, see attached and/or see http://www.loc.gov/preserv/digital/dp-refguidelines.html and http://www.bampfa.berkeley.edu/moac/
- While standards established in the late 1990s at the Library of Congress and the National Archives proposed 300 ppi as the minimum standard for digital imaging, 600 ppi is the current standard for digital imaging, e.g. Museums and the Online Archive of California (MOAC).

MOAC Standards

	Resolution	File Format	Color Target
Master	> 600 ppi or 3000 pixels / longest (whichever possible / yields bigger file size) OR 20 MB target	TIFF (uncompressed)	Yes
Submaster	< 3000 pixels / longest dimension; color bar cropped out	Tiff (uncompressed)	No
Access	> 150 pixels / longest dimension (recommended 640, 800 or 1024)	JPEG (compressed)	Optional (recommended No)
Thumbnail	= 150 pixels / longest dimension	JPEG (compressed)	Optional (recommended No)

- The CHM best practice and long term goal is to meet the most current digital imaging standards as outlined above. Any future CHM digitization efforts and equipment purchase will make every effort to be consistent with these standards. Further testing and research is necessary, however, to determine whether these standards (600 ppi) are appropriate when scanning text only documents for legibility, in such cases, 300 ppi may be appropriate.
- The CHM has also established a minimum digital imaging standard of 300 ppi order to make immediate use of existing digital assets created internally or provided by external contributors.

Summary of Digitization Standards:

Image Description	CHM Minimum (Interim)	CHM Best Practice
Alphanumeric text	300 ppi	300 ppi
Handwritten text	300 ppi	600 ppi
Schematic Drawings	300 ppi	600 ppi
Illustrations	300 ppi	600 ppi
Photographs	300 ppi	600 ppi

N.B. More detailed scanning standards to be developed under a separate document, e.g. bit depth

9. Surrogate Image Storage: File Folder Structure and File Naming Conventions

- Although the CHM best practice is to assign accession numbers and create an individual record (e.g. Database Metadata) in the Collections Database for each item in the CHM Collection, surrogate images (e.g. scans of documents, etc.) can also be saved using file folder and file naming conventions prior to receiving individual records in the Collections Database.

File Folders Conventions:

- All surrogate images will be stored on a permanent server location in an **Resource Type Folder** according to its related Collection or resource type. Storing images by resource type at the top level (as opposed to topic), will allow CHM to manage data by collection type, therefore catering to the different handling and delivery needs of each media type. Also users will always be able browse the entire file folder structure by topic using a Google search engine.
  - 🗁 still_image →photograph collection (tiff files)
  - 🗁 moving_image →video collection (dv files)
  - 🗁 object →reference photos of artifacts (jpg files)
  - 🗁 text →document collection (pdf files)
  - 🗁 software →software collection (exe. files and other)

- Each Resource Type Folder will contain a **Name Folder** of specific items organized alphabetically by Manufacturer, Software, Surname or Event Name/Year, as "intuitively" appropriate, e.g. :
  - 🗁 still_image
    - 🗀 abc
    - 🗀 aiken
    - 🗀 altair
    - 🗀 amdahl

- Each Image Name Folder may contain a further **Name Sub-Folder** if the file contents of that folder exceed fifty files (TBD), e.g.:
  - 🗁 still_image
    - 🗀 dec

- pdp-1
- pdp-6

- If a file folder name has more than one term “Hewlett Packard”, use “_” to separate the terms, e.g. hewlett_packard
- All elements will be in lower case.

File Naming Conventions:

- In order to ensure added file identification in case a file is disconnected from the Collections Database as well as to facilitate free text searching of files, the surrogate image file will also contain metadata.
- Although each collection will require a slightly different approach for file naming, all file names should use a “.” between metadata elements and a “_” within metadata elements. It is allowed to use other characters within metadata fields if they are considered integral to the object name or title, e.g. the hyphen in “PDP-1”. N.B. Windows will not permit the use of any of the following: / \ : * ? | " < >
- All elements will be in lower case
- If a metadata element is not known or not relevant to the item, this field can be bypassed.
- The following file name structure should be followed as closely as possible for each collection:

  Physical Objects:
  [manufacturer][machine or object name][accession #].ext

  examples:
  dec.pdp-1_printer.X126-84.jpg
  keuffel_&_esser.telegraph_sending_and_receiving_device.102626504.jpg

  - Replace old accession number “.” with “-”, e.g. X126-94
  - To capture multiple views of the same object, add a number to the file name: apple.macintosh_keyboard.X126-84.1.src.jpg
  - To capture a system view of an object {keyboard with full PC}, use an “S”: apple.macintosh_keyboard.X126-84.S.src.jpg

  Image Collection:
  [manufacturer][machine or person or event][date][accession #][format].ext

  examples:
  dec.pdp-1.1964.102525403.src.jpg {photo of machine}
  dec.pdp-1_bell_kotek.1964.102525404.src.jpg {photo of machine with people}
  cobol_reunion.1984.103434504.src.jpg {event, too many people to list}

  - [format] refers to image size: sm, lg or src

Text Collection:
[manufacturer][machine][title][date][accession #].ext

example:
dec.pdp-1.principals_of_operation.1964.102525403.pdf

- Although as a rule, surrogate images will be assigned accession numbers before being stored in the File Folder Structure, Collections that currently include large quantities of previously scanned surrogate images (such as the Image Collection), can be stored in the File Folder Structure prior to obtaining accession numbers. These items will receive file names with descriptive information (as above) and accession numbers will be added at a later date as these items are entered into the Collection Database, e.g.
  - apple.macintosh.1964.src.tiff
  - chm_fellows.2003.src.tiff
  - amdahl.1996.src.tiff
  - cobol_reunion.1984.src.tiff

- Many surrogate images will also have multiple formats or sizes for a variety of uses (e.g. thumbnail for web use, large for internal use, etc.), which will also be stored in the above file folder structure. In order to identify their size/use we will adopt the following naming convention:

	Resolution	File Format	File Name e.g.
Source	< 3000 pixels/longest dimension	TIFF (uncompressed)	apple.102525401.src.tiff
Large	> 150 pixels/longest dimension (640, 800 or 1024)	JPEG (compressed)	apple.102525401.lg.jpg
Small	= 150 pixels/longest dimension	JPEG (compressed	apple.102525401.src.jpg

- Since the CHM will make its File Folder Structure public, due to security reasons it will be necessary to protect the “src” files under a separate folder in the Image Folder or Sub Folder, thus allowing the capability for machine instructions to protect all files under “src”:
  - still_image
    - dec
      - pdp-1
        - dec.pdp-1.102626401.sm.jpg
        - dec.pdp-1.102626401.lg.jpg
        - src
          - dec.pdp-1.102626401.src.jpg

10. CHM Collection/Digitization and Metadata Priorities

- The Text and Photograph elements of the CHM Collection are ideal candidates for digital imaging. The chart below breaks these elements of the Collection into document types and assigns scanning priority according current perceived demand for access. The CHM digitization project will focus initially on digitizing photographs and a subset of manuals.

Digital Imaging Priority

CHM Text Collection	Percent of Collection	Scanning Priority
Manuals	20%	1
Schematics	15%	2
Brochures	15%	2
Photographs	15%	1
Technical Reports	15%	3
Serials	15%	2
Company Corresp.	4%	2
Books	1%	3

11. Born Digital Donations

- The CHM Cyber Collection will also serve as the repository for "born digital" donations to the CHM Collection. Such donations include items whose primary source exist in digital form, e.g. online publications.
- The CHM will process born digital donations using established registration procedures and database standards established earlier in this document:
  a. Offer to donate is sent to Registrar
  b. CHM Staff reviews offer and accepts/rejects
  c. If accepted, Registrar creates Deed of Gift
  d. Donor sends donation "materials" via email or on CD
  e. Donor signs Deed of Gift
  f. Registrar assigns Accession Number
  g. Relevant CHM staff member (e.g. collections specific) enters minimal metadata for each individual item and saves digital files according to file naming convention outlined above.

12. Outstanding Issues

- Copyright/IP policy regarding making digital versions of CHM Collection available online
- Software collection digitization steps

Text and Image Collections – Work Flow