1. (16 points) A printed text document can be converted to digital formats by a choice of methods:

(i) digitization by scanning
(ii) digitization plus optical character recognition
(iii) retyping with SGML markup

(a) (12 points) What are the advantages and disadvantages of each of these three methods? You may use phrases or bullets to give your answers instead of complete sentences

	Advantages	*Disadvantages*
Scanning	*- fastest and least human resource intensive of the alternatives* *- captures page exactly as is, important for image data*	*- may destroy some originals* *- scanning errors require manual correction* *- most difficult to search in comparison to all other alternatives* *- most space required for a usable copy*
Scanning + OCR	*- allows access to textual data for searching* *- compact representation for storage*	*- errors in OCR (especially for c documents or ones with difficult fonts or handwriting)* *- correction of bad OCR comparable to manual input cost*
SGML markup	*- most clean means of digitization* *- adds structure to the data* *- can encode most formatting and layout variations (with appropriate standard)* *- flexibility in granularity of markup*	*- manual labor cost* *- slowest of alternatives* *- specialist cost in coding markup* *- potentially more difficult to use for non-specialist than plain raw text.* *- may still require scanning to embed images*

(b) (4 points) Under what circumstances would a user be unsatisfied with all three digital manifestations and want to use the original printed copy?

Assuming a user has access and rights to digital versions, he may still want to consult the original if:
- the user is interested in aspects of the book other than the textual material and if a page scan of the book does not provide the relevant information (e.g., binding information, page texture);
- the book is not amendable to scanning (e.g., too fragile) or the scanned copy may be too poor for the specific use. Today's planetary scanners have mostly alleviated these problems;
- the authenticity of the document is important or if the original has sentimental value;
- the portability and easy access to the resource is important.

2. (4 points) We saw that CCITT includes codes for run lengths of zero. We know that lines in CCITT can have a black run of zero length because the standard assumes the start of a line is a white run. What is the black run of zero used for?

The makeup codes specify a run of pixels of a multiple of 64 pixels in length. Terminating codes in CCITT signify a run exactly n pixels in length where n is between 0-63 pixels in length. CCITT uses a a makeup code appended with a terminating code to encode an arbitrary length run.

A black run of zero is a terminating code and is used together with the makeup black codes to signify a length is exactly a multiple of 64 in length. For example if a 128-pixel run of black was seen it would be encoded as 2x64 black makeup code, followed by a 0 length black terminating code.

3. (6 points) Given the query "$word_0$ $word_1$ $word_2$", construct two example documents d_1 and d_2 that would be ranked differently by the Euclidean and cosine similarity calculation methods. Show your calculations.

Let d_1 be a document with three words: $word_0$ $word_1$ $word_2$.

Let d_2 be a document with six words: $word_0$ $word_1$ $word_2$ $word_0$ $word_1$ $word_2$.

We make a simplifying assumption that the word weights have all been set to one. Then q is equivalent to d_1, as they have the same ratio of words. But the frequency of the words are different in d_1 and d_2, so their Euclidean distance differs. For completeness we show this below, using the abbreviation w_n *for* $word_n$*:*

$$Sim_{Euclidean}(q,d_1) = 1 / sqrt((abs(w_{q0}-w_{d0})^2 + abs(w_{q1}-w_{d1})^2 + abs(w_{q2}-w_{d2})^2))$$
$$= 1 / sqrt((abs(1-1)^2 + abs(1-1)^2 + abs(1-1)^2)) = 1 / 0 = undefined\ (identical)$$

$$Sim_{Cosine}(q,d_1) = 1/(w_{d0} + w_{d1} + w_{d2}) * ((w_{q0}*w_{d0}) + (w_{q1}*w_{d1}) + (w_{q2}*w_{d2}))$$
$$= 1/(1+1+1) * (1+1+1) = 1\ (identical)$$

But this is not the same for d_2:

$$Sim_{Euclidean}(q,d_2) = 1 / sqrt((abs(w_{q0}-w_{d0})^2 + abs(w_{q1}-w_{d1})^2 + abs(w_{q2}-w_{d2})^2))$$
$$= 1 / sqrt((abs(1-2)^2 + abs(1-2)^2 + abs(1-2)^2)) = 1 / sqrt(3)$$

$$Sim_{Cosine}(q,d_2) = 1/(w_{d0} + w_{d1} + w_{d2}) * ((w_{q0}*w_{d0}) + (w_{q1}*w_{d1}) + (w_{q2}*w_{d2}))$$
$$= 1/(2+2+2) * (2+2+2) = 1\ (identical)$$

4. (16 points) Metadata and other annotations to a document can be stored either in-place (as part of the document) or in a separate file (standoff annotation). Baca (1998) terms this distinction as the *source* attribute of metadata (internal or external)

i) (4 points) Give an example of each: internal metadata and external metadata.

We've learned several examples of each. Internal formats for metadata include: HTML and TEI headers, MIME type extensions, graphic headers such as ones used by GIF, JPEG and MPEG.

A good example of an external format is the MARC standard for library materials. Other examples include standalone abstracts of papers, external document type definitions and style sheets in XML, and external metadata packages referenced by a URI in the Warwick framework.

ii) (9 points) Discuss how the values of other metadata attributes might influence a designer's decision to use external or internal metadata.

The nature, status, semantics, creation and level attributes of metadata can all influenced the choice of where to place the metadata externally or internally.

Lay metadata that is used by an individual on a small scale is more likely to be stored with the digital item itself. Expert metadata may be voluminous and be used to serve as an access point for the item, and be more likely to be stored externally.

Similarly, collection level information cannot be associated with an item itself so would have to be stored externally. Item or subitem information may be stored within the item itself.

Dynamic metadata, such as transaction logs, would normally be stored external to the document itself. Likewise, manually created metadata (e.g., annotations and comments) may be subjective and more valuable than the digital document itself, and may be stored external to the document.

iii) (3 points) Use your answer in ii) to explain why designers chose to use the source they did both examples in i). Be brief.

HTML and TEI headers are internal as indexing utilities process this information to build indices for their web page and literary document collections. MIME type extensions and graphic headers are also internal as they contain descriptive metadata needed to decode and display the object.

MARC records and paper abstracts are external as they are used as part of a larger mechanism in indexing to help searcher locate relevant library records. External document type definitions and style sheets in XML help to separate the validation and rendering functions of an XML structured document from the data itself. External metadata packages in the Warwick framework can be used for many different reason. One is to point to common metadata attribute and values for certain records (e.g., intellectual property records) and reduce the redundancy in records.

5. Briefly define **four** of the five terms (8 points):

Companding:

Compaction + Expansion: In general, the use of both compaction and expansion in the representation of a scale. In audio coding, the use of more sampling for low frequency audio than high frequency ones, as the human auditory system is more sensitive to the former.

Dublin Core:

A simple metadata standard comprised of fifteen fields. The fields can take on arbitrary values, but the standard suggests that values take the form of other recommended standards (e.g., dates to be standardized in accordance to an ISO standard). Specifically geared towards interoperability of metadata across different standards and fields.

CCITT Fax IV:

A standard for bi-level facsimile transmission. It is based on CCITT Fax Class III standard. The Fax IV standard is used on digital transmission networks in which error detection and correction is handled externally (i.e., assumed to be taken care of by earlier stages of the processing pipeline). It achieves better compression ratios compared to Fax Class III by comparison. It has provisions for greyscale and color encoding as well.

Raster image:

An image represented as a grid of pixels. Each pixel contains brightness and color information, and the image as a whole may specify resolution in terms of pixels per unit length. In contrast to vector images, raster images often lose clarity when scaled or rotated.

Word inflection:

Variation of a word as changed by tense, gender, case, or number. For example, the word jumps *is the present tense inflected form of the base word* jump. *Some querying and indexing systems remove word inflection to save space in their indices.*

End of Paper