Xerox copies to:
Glaser
Saltzer
Graham
Evans

June 2, 1966

M. D. McIlroy

MESSRS. F. J. CORBATO
E. R. VANCE
V. A. VYSSOTSKY

Canonical character strings as proposed for Multics in
BC.2.02 will cause some severe difficulties for computing
and even for mere editing. Herewith are some particular
troubles:

1. Simple bulk character I-$\emptyset$ will be impeded, since
reformatting will be necessary on every line. The
fact that each card reader and each printer on Multics
will do more I-$\emptyset$ than all the typewriters put together
makes this an important issue.

So our standard
character form should be
whatever the printer
needs ? ? ?

2. Canonical strings are not good data for computation,
since string operations, notably substring extraction
and concatenation, do not preserve canonical form.
String scanning and comparison become a complicated
matter. For unfortunate examples, consider the PL/I
statements,

$\qquad$ i = length("␣␣␣␣");

$\qquad$ if ("␣"ll"␣␣") = "␣␣␣" then...

$\qquad$ a = substr ("␣␣␣␣", 1, 1);

$\qquad$ i = index ("␣␣␣␣", "␣␣")

If length, etc are
represented right
this shouldn't make any difference

3. Information is lost in canonizing. Removing tabs loses
alignment information. Stripping trailing blanks may
lose meaningful character positions.

nonsense

4. Though canonizing will often make two streams that look
the same read the same, it does not make two substreams
that look the same read the same, and it is substreams
that we are usually interested in comparing. The
discipline on backspace and new line does serve this
purpose, the discipline on tabs and special vertical
feeds does not.

requires
appropriate
substring extraction
routines to make
it work.

5. Saving space in bulk storage is another, unarticulated,
purpose of the proposal. Experience with crunch decks
shows that, though the saving due to suppressing trailing
blanks is substantial, the saving due to the relative tabs
is unimportant.

agreed

relative to full card images, yes.
" " compressed " " not so true.

6. Few of our output devices have up feeds, so text
   including these will usually be unprintable. Moreover,
   text involving other than initial down feeds does not
   satisfy the criterion that streams which look the same
   read the same.

7. Identical overstruck characters are useful - this makes
   boldface on a scope, or perhaps rings the bell insistently
   on a teletype - they should not be squeezed out.

Recommendations:

1. Scrap the relative horizontal tab.

2. Make right margin control possible. Consider keeping
   one trailing blank when there is no right margin.

3. Vertical feeds other than down feeds at the beginning of
   line are too irregular and specialized for canonizing.

4. Overstrikes if we want them at all should be general,
   at the programmer's discretion, not reduced to one of
   each character.

5. For people who wish to use tabs, tabs should be permitted
   in canonical text. The result of backspacing across a
   tab is to suppress canonization. The possibility of
   replacing tabs with spacing to known tab stops as described
   in the proposal should also be available.

MH-1371-MDM-gam   M. D. McILROY