**What makes this digital presentation of Daily Worker 1924 - 1936 superior to others?**

**Statement by Marty Goodman MD of the Riazanov Library digital archive project, who either physically scanned from paper or downloaded from one or another web page most of the pages presented here, and who laboriously assembled into issue files and constructed names for each such issue file, for each of the roughly 5000 issues... roughly 30,000 pages... of *The Daily Worker* digitally presented here (with some help from Robin Palmer).**

There are several other presentations of Daily Worker one can access on line as we enter the final stages of preparing this one for release to the public.

The Library of Congress' "Chronicling America" web page contains a large fraction of what we have here. In fact, it is from there that we GOT well over half of the material presented here.

Some other sites are now also offering the material the Library of Congress presents, on their web pages.

Those universities which pay Proquest for access to their digital holdings provide at terminals at that university access to what is said to be an digital archive of *The Daily Worker* from the first issue on January 13, 1924, to the last issue printed of the paper on January 13, 1958.

Both at the time of my writing this allow downloading of the material there.

What makes this digital archive different... and (I would argue) superior to any other currently available (at least in the target period of 1924 - 1936) ??

(1) We present the paper as multi-page pdf files of each issue. Library of Congress and Proquest present it as page at a time display and download. Requiring arduous multiple clicks / keystrokes to download each page.

In the case of Proquest, a commercial operation that seeks to own (or act as if it owns) the material it offers for sale, it seems plausible that the choice to offer this material one page at a time, with multiple clicks and operations required to download each page, is intentional.

In the case of the Library of Congress, and other academic sites that offer the Library of Congress archive of Daily Worker, I've some reason to believe this page at a time download approach to presenting the material is done due to lack of insight and poor choices on the part of those setting up the presentation (and perhaps also hesitation given the immense amount of extra work it is to carefully construct multi-page issue files), not out of malice or active desire to thwart downloading the material.

I say this because my personal experience with the Library of Congress, when I sought to obtain their files for us to present here, was that of immediate, kind, and generous cooperation in assisting me with setting up for automatic mass downloading from their site Also, people at the Library of Congress on several occasions made very clear to me their DESIRE was make their information as freely and widely available to all, as possible, and that they welcomed their material being re-distributed by others, on other web pages (with proper acknowledgement, of course, of their role in providing the material). My experience of the Library of Congress has been that of seeing their actions as those of principled librarians who wish to serve the public. Not jealous petty shop-keepers, like the folks at Proquest. Or at Yale, for that matter, tho that is another story I will not go into here.

Note that the way the material is presented on line by Library of Congress web site (and by us at our Marxists.org web site) it's relatively easy for a sophisticated computer type to effect automatic mass downloading of entire archives. Proquest has its material organized so as to maximally thwart such efforts.

Our presentation of this (and other newspaper / periodical archives on Marxists Internet Archive is alwaysas multipage issues one can download with a single click. Try as I might, can see NO justification for presenting this sort of material as multiple-click-requiring page-at-a-time downloads.

2. We assign to these full issue pdf files intelligently constructed, highly functional file names. Both the file names and the notation for the files on the web page where one goes to access the files offer the Volume and Issue number and the date of each issue. In the file name, this information is provided in a fashion that both forces chronological ordering of the files if they are placed in given folder and displayed alphabetically **and** which makes them easily and rapidly **human-readable**, because in our file names we present the month as a three letter abbreviation and separate individual units of data (volume, number, issue type, day, date, and year) with hyphens, in contrast to what I regard as the inept and incompetent approach that is the norm in much of academia of assigning file names consisting entirely of numbers run together, and usually (as in the case of material offered by the Library of Congress and by Proquest) without volume and issue information in the file name for the issue or page.

In the case of this *The Daily Worker* archive specifically, sometimes there is a Sunday issue in the run, and sometimes there is not. Very occasionally a day where normally one would expect an issue there was none ever printed, for one reason or another. Without providing both dates and issue numbers, one has no way to tell whether what appears to be a missing issue is one that never was printed, or one that was printed but currently lacking from the collection being presented. And one is far less able to identify anomalies in the printing schedule.

One person centrally involved in the creation of the Library of Congress digital archive of *The Daily Worker*, when I brought this up to that individual, responded "Issues numbers are not always reliable... they are sometimes given in error." Yes, that is literally true. But what he neglected to observe is that such anomalies occur in this archive at most two or three times per year (per 330 or so issues... thus about once per 100 issues). So the vast majority of the time issue numbers DO serve to accurately check what's going on. Further, when such errors occur, they are nearly always easy to identify. We've specifically identified most of them in our presentation of *The Daily Worker*, noting where an issue bearing one issue number really should bear a different one, or where an issue number is accidentally repeated for two successive issues. This person's observation in no way changes the fact that providing / paying attention to issue numbers critically aids in identifying the integrity of and maintaining (improving) digital (and other) archives of periodicals.

3. One rather unusual (and arguably perhaps just a wee bit "controversial") thing I've instituted in the creation of file names in those digital archives of periodicals that I've created over the last dozen years is the occasional, infrequent inclusion of comments ... in effect tiny bits of "metadata"...in the file namefor a given issue. Thus, if an issue bears an obviously incorrect issue number, I note this **in the file name** of that issue. A few other anomalies, which can be expressed in a very few short words, I sometimes also note **in the filename**. For example, if the issue is missing one or more pages, or if pages in the digital image have significant damaged, such as holes in the paper or a distorted scan.

It's a somewhat subjective decision on my part, personally, what "micro meta data" I choose to put in a filename, vs when I decide the information is either too long or not important enough to stuff into the actual file name of the issue. But my objective criteria are to use this approach where I see it as greatly facilitating my ability to maintain and improve the archive in the future, allowing me to easily spot issues that would benefit from being upgraded / replaced particularly, just by looking at a directory of the file names scrolling by.

One must be careful not to use this technique too often, or to bloat the length of file names too much when one does use it. But I believe, used judiciously, this (a bit odd on the face of it) technique can aid in maintenance and future improvement of digital archives of periodicals.

4. While one can debate how much superior the above aspects of this archive make it, compared to others, there can objectively be zero doubt that the point I am about to make makes this digital archive uniquely and **very significantly** superior to **all other** offered digital archives of "*The Daily Worker*".

This is **the only** available digital archive where each issue is, with a flag in its file name, identified as to whether it is a New York or a National edition during those years (1925 - 1936) when two different editions of the paper (often with different content) were printed each day. And where those issues are presented on web pages that inform regarding which edition type issues are being offered.

Why? Simple: Because until I created this digital archive, **no** institution that either maintained archives (microfilm and/or paper) of *The Daily Worker*... and **no** institution presenting this material digitally on line... was even AWARE there were two issue types. In my judgment, this is inexcusable, and due to a high degree of laziness and slovenliness in their curation of their archives, for all you have to do is look at a number of issues of the paper carefully to quickly realize that this is going on. After all, this is not an obscure or unimportant paper. It's one that's very often referenced by those addressing the history of the communist movement in the USA in the first half of the 20ieth century.

The fact no library or special collection knew about the existence of two different issues of *The Daily Worker* published each day [not even Tamiment Library, which holds the Communist Party USA's own archives] also meant that no catalog listing of contents of paper, microfilm, or digital archives being offered in the past provided any clue as to what National vs New York editions were in the holding. Which, in turn, presented for me immense problems in trying to locate and obtain copies of both New York and National editions of the paper in those years (1925 - 1936) in which such were printed. Which is why we currently can offer both the New York and National editions only for some of the years both were published, and why some of our runs of one or the other edition in the years we present both are pitifully fragmentary.

5. The following is a sophisticated technical point, requiring some knowledge of computer file format technology. But it's not, I believe, an entirely insignificant one:

We on Marxists.org present all our digital form archives using generic ordinary pdf format.

Many institutions... including the Library of Congress, Archive.org, and Proquest... present digitally their pages of *The Daily Worker* in pdf/A format, not generic ordinary pdf format.

Administrators of digital presentation of archival material somewhere along the line saw that the "A" in "pdf/A" stood for "archival", and saw one or another technical paper (which they no doubt did not understand) and ignorantly concluded: "Archival scans of paper represented by pdf files should be presented in pdf/A file format". They may also have in the articles they read about pdf/A format repeatedly encountered references to "pdf/A format preventing alteration", and which further convinced these computer-ignorant administrators THIS is how archival files should be presented.

The facts are these:

Pdf/A format was created SOLELY for use with scale-able font / vector graphic type pdf files. NOT for pdf files that consist entirely of bit map, photographic-type images of pages, or pdf files that consist of bit map images of text that also have embedded OCR (Optical Character Recognition). It was created to **gently remind** [**not to "enforce" !**] people in the legal and other document-production professions to not distort the pdf file by removing from the pdf file in question the library of fonts for the scale-able fonts used in the document, so that when it is printed out or displayed on the screen it will look perfect and proper.

There is zero positive value in employing pdf/A format for the pure bit map-type (with no scale-able fonts or vector graphics) file output of a scanner or other camera-like system imaging paper or microfilm. **ZERO**.

It is misguided folly to believe pdf/A **prevents** someone from altering a given pdf file. While it is a minor hassle, and an extra bit of work, to remove the pdf/A restrictions and turn the file into an ordinary pdf file, this can be done with any of dozens of pieces of software, **including** the universally available pdf display program, downloadable free of any charge, and found on a large fraction of existing computers, Acrobat Reader. It also happens that if one simply combines individual page pdf/A files into a multi-page pdf document using Acrobat Pro, the pdf/A restrictions go away, and the result is a normal generic pdf file. That's how I purged pdf/A restrictions from the multi-page issue files I made from the individual pdf/A page files originally offered by the Library of Congress,in, and Archive.org.

But so long as a file is IN pdf/A format, you cannot clip out of it a portion of the page. You have to spend the time to convert it back into an ordinary generic pdf file first. An annoyance.

Being able to digitally clip out a column from an image of a newspaper can be absolutely essential and critical to using the embedded OCR in the image of that file, for often OCR put into such files fails to recognize column boundaries, and the only way to copy text created by that OCR is to first CROP OUT each individual column!

In general, serious professional use of archival digital material by students, researchers, scholars, and historians repeatedly entails clipping out a portion of the page presented, in order to cite some specific article or portion of article.

Having the page in pdf/A format accomplishes **NO** positive purpose, but forces an irritation... an extra step (the removal of pdf/A restrictions) ... before one can properly use the file.

Finally (and I admit this is a partly an academic point, but it is not entirely without merit) ...

Generic pdf file format has for many years been placed in the public domain. You can look up the details of the structure of that format in many places, and freely legally share that information with anyone you wish, including re-posting it wherever you like.

But as of my writing this little essay, pdf/A format remains a PROPRIETARY format. True, documentation of it can (with some difficulty) be obtained from one or two sites. But technically it is illegal to give someone else a copy of the technical details of pdf/A file format.

This makes it further unsuitable for archival use.

If you are digitally archiving images of documents, it is absolutely essential that the specifications for the file format you are using be in the public domain and very widely available, in plain English (or whatever other language you speak). So that way, decades or centuries from now, if the systems for displaying and printing those files are gone, you can from the information on the file format program existing systems to display and print such ancient archival files.

[When reading the above about pdf/A, please note I've been programming computers for near 60 years. When I was 13 I was programming IBM 7094 mainframe computers at Columbia University in assembly language. I for a while wrote a regular column in a computer magazine. For this essay, I consulted with one of the world's greatest experts on computer operating system directory structure, Fred Cisin, who created "Xenocopy", which allowed porting one of many hundreds of operating system type disks to and from IBM PC type format. I specifically consulted with him regarding pdf/A format, and as far as he could tell, what I've said about it is accurate.

If any reading this find I have made a significant mistake in what I've written, above, concerning pdf/A file format... a mistake that would change my central point that using pdf/A file format for archival bit map scans of paper and microfilm is inappropriate... a mistake in what I claimed, or a failure to note critically important other relevant information... **PLEASE LET ME KNOW.** I will change the content of this document immediately, if I can verify what is offered to me.]

=======================================================