Telecommunications Industry Forum

**Sponsored by the Alliance for
Telecommunications Industry Solutions**

Comparison of HTML, PDF, and SGML

TCIF Memorandum
TCIF-IPI-97-005
Issue 1, Revision 1, 10/24/97

Comparison of HTML, PDF, and SGML

Prepared by the TCIF Information Products Interchange Subcommittee, http://www.atis.org/atis/tcif/.

To order, please contact TCIF, (202) 434-8844, FAX (202) 393-5453.

For more information, please contact Donald F. Pratt, (732) 699-4012, dfp@ims.bellcore.com.

TradeMark Acknowledgments

Adobe, Acrobat, and PostScript are registered trademarks and Acrobat Exchange and FrameMaker+SGML are trademarks of Adobe Systems Incorporated.

ADEPT Publisher is a trademark of ArborText, Inc.

Macintosh is a trademark of Apple Computer Incorporated.

Microsoft is a registered trademark and Windows is a trademark of Microsoft Corporation.

Netscape, Netscape Navigator, and the Netscape N logo are registered trademarks of Netscape Communications Corporation.

SoftQuad Panorama PRO is a trademark of SoftQuad Inc.

UNIX is a registered trademark of X/Open Company, Ltd.

1 Introduction

1.1 Purpose and Scope

This document provides a brief comparison of three formats for electronic documentation, concentrating on differences that may be meaningful to the technical support staffs of telephone service providers who might be using documentation provided in these formats by telecom equipment vendors.

A comparison of the uses and usefulness of these formats must take into account the tools ("browsers") that present them to the user. These tools may differ in ways that have nothing to do with the inherent differences between formats, but the differences are nevertheless real if the choice of tool is constrained by the format and reasonable consideration of cost. We have distinguished those differences that appear to be inherent in the format from those that are most likely temporary differences in the feature sets of the browsers. Expect some "temporary" differences to persist through the next few versions of the browsers.

The formats compared are the three most widely used formats (other than plain text) for electronic documents: PDF (Portable Document Format), viewed with Adobe® Acrobat® 3.0 (no cost for Reader, $295 for the package that includes Exchange™, which is needed to create hyperlinks and annotations); HTML (HyperText Markup Language), viewed with Netscape® Navigator® 4.0 beta (free for evaluation, $59 later); and SGML (Standard Generalized Markup Language), specifically TIM (Telecommunications Interchange Markup), viewed with SoftQuad Panorama PRO™ 2.0 ($49; $195 with an interactive stylesheet editor). Go to these sites to download the free or evaluation copies:

- **www.adobe.com/prodindex/acrobat/**
- **home.netscape.com/comprod/products/communicator/**
- **www.sq.com/products/panorama/**

1.2 Acknowledgments

Some of the content of this memorandum has appeared previously in unpublished documents written by Eve Perris and Donald Pratt of Bellcore and David White of ArborText, and there has been input from many others.

1.3 Use of product names

It is contrary to the intent of this document to provide any support whatsoever for particular products of, or practices favored by, some manufacturers and suppliers over others. Names of commercial products are used in this document for illustrative purposes only. Anything that can be construed as an evaluative statement is intended to refer to the document formats, not the tools used to

view them. No endorsement or evaluative judgment of any product, service, or provider is intended, and none should be inferred.

2 Differences for End Users

Appearance: PDF files (Figure 4-2) look like the printed document (not always exactly: unavailable fonts and EPS graphics can cause problems). SGML files (Figure 4-3) may look *somewhat* like the printed doument, but appearance is controlled by a user-selectable stylesheet, so it may be different for every user. HTML files (Figure 4-1) can have stylesheets, but generally styles are controlled by the provider rather than the end user, who can change only the basic font and color.

HTML and SGML viewers wrap text to fit the viewing window, so horizontal scrolling is unnecessary. Panorama (SGML) also sizes graphics to fit the window, with a quick way to enlarge them. There is no resizing of graphics in Netscape. **Advantage: PDF or SGML**, depending on whether you value faithfulness to the original over flexibility and legibility.

Text search: PDF is a binary file format, and searching for text is inherently slower than searching in HTML or SGML, which are ASCII. On the other hand, because they are ASCII, HTML and SGML incorporate graphics (and anything else other than text) only as external files. Text searches bypass any text in the graphics; in a TOP (Task-Oriented Practice) document that is mostly flowcharts, it may be hard to find *anything*. Acrobat will find words in drawings (but not scanned photos or screen captures), because text in drawings is still text.

Panorama has two search features the other viewers don't have. One, which could potentially be part of any viewer, is an occurrence density display that shows graphically along the vertical scroll bar every occurrence of the searched-for string. The other, which is an inherent advantage of SGML, is a word-in-context search: to limit the hits to those that may be most relevant, you can search only for occurrences of the word in certain elements, such as titles or requirements. (None of the tools provide a way to limit searches to part of the document, such as an introductory section.) **Advantage: PDF or SGML**, depending on whether you need to search for text in graphics.

Navigation and annotation: Text search (above) is a way to find what you want when you don't know where it is; navigation aids help you get to a place you know you want to be. Many people have used the bookmark feature of Netscape, which lets you return to an important document with one or two mouse clicks. But this is the only navigation aid Netscape provides. Netscape also provides no way to make annotations, short of editing the document source file and saving it as a new copy.

In Panorama, you can bookmark any specific place in a document. You can also create your own hyperlink between any place in one document and any place in another, for instance between a section in a vendor's switch documentation and one in your own company's technical bulletin. You can also add an annotation to

any place in a document. Your links and annotations appear as icons in the document, but the original document is not modified: the links and annotations are saved in a personal "Web" file. Acrobat allows similar links and annotations, but it can create them only in Exchange, part of the $295 commercial version, and its links and annotations become part of the file, so the user must have permission to edit the file or must save a personal copy, and if a new version of the document is delivered, the links and annotations must be re-created. (Panorama links may or may not survive updates, depending on whether the anchor's ID or position, whichever was used in the original link, is still the same.)

A unique Panorama feature made possible by the structural information in SGML is automatic generation of tables of contents, including lists of figures, tables, or any other element type. **Advantage: SGML**, PDF second, HTML far behind.

No difference: All three formats allow multimedia objects (movies and sound files) to be included as separate, linked files, and the viewers can be configured to launch the application that views or plays them. (They can also launch each other, but their links to documents of other types cannot point to specific places within those documents, the way Panorama links between SGML documents and Acrobat links between PDF documents can.) All three viewers allow users to fill out and submit forms. For all three formats, add-ons to the basic browser products can create full-text indexes to speed up searches.

3 Differences for Originators and System Administrators

Creating the format from current applications and legacy files: Any Macintosh™ or Windows™ application and almost any UNIX® application can create a PostScript® file through the Print menu, and almost any PostScript file can be turned into PDF. Any legacy file that can be opened by a current application can also be turned into PDF. There is nothing so automatic about creating HTML or SGML. Several commercial products are available to help create HTML, which is a fairly simple format, but to create SGML there has to be information about structure that is not normally included in a word-processor or desktop-publishing file. Conversion to SGML generally requires expensive manual cleanup. The most cost-effective strategy for routinely creating SGML is to switch to an SGML-aware authoring applications like ArborText ADEPT Publisher™ or FrameMaker+SGML™. SGML authoring applications are expensive and there is still cleanup, but only once per document, not every time the doument is issued. HTML can be created automatically from SGML. **Advantage: PDF**, HTML second.

Revising and reusing: PDF files are not completely static, but text is stored line by line, and tools that exist for editing PDF can edit only one line at a time. Typos can be corrected this way, but major rewording is impractical. There is also no practical possibility of converting a PDF file back into a word-processor or desktop-publishing file, although text passages can be extracted and most graphics can be retrieved. With HTML and SGML, text and graphics can be edited directly or reimported to authoring applications (reimporting requires application-specific filters). HTML files may not contain rich enough or consistent

Differences for Originators and System Administrators
3
TCIF Memorandum
Comparison of HTML, PDF, and SGML

enough markup to allow all formatting features to be re-created, but imported SGML can come very close to the original with very little manual cleanup. **Advantage: SGML**, with HTML second.

Installing the viewer: Users can install Acrobat or Netscape themselves and be ready to use PDF or HTML files. Panorama, out of the box, does not know about TIM (Telecommunications Interchange Markup) or most other versions of SGML, so several extra files need to be installed with it, and initialization files have to be edited. These can be packaged for a custom installation, but that will require some effort by a system administrator.

Panorama comes with a stylesheet for viewing HTML (since HTML is a variant of SGML), but this does not make a Web browser unnecessary. Panorama will complain about any violation of structural rules (in our experience, most HTML files violate the structural rules in the SGML definition of HTML). Netscape ignores the structural rules and does the best it can to display any HTML file. **Advantage: PDF and HTML.**

4 Sample Screens

These screen samples are taken from a Windows 3.1 800x600-pixel screen with 256 colors. (If you are looking at this document in Acrobat, set magnification to 200%.)
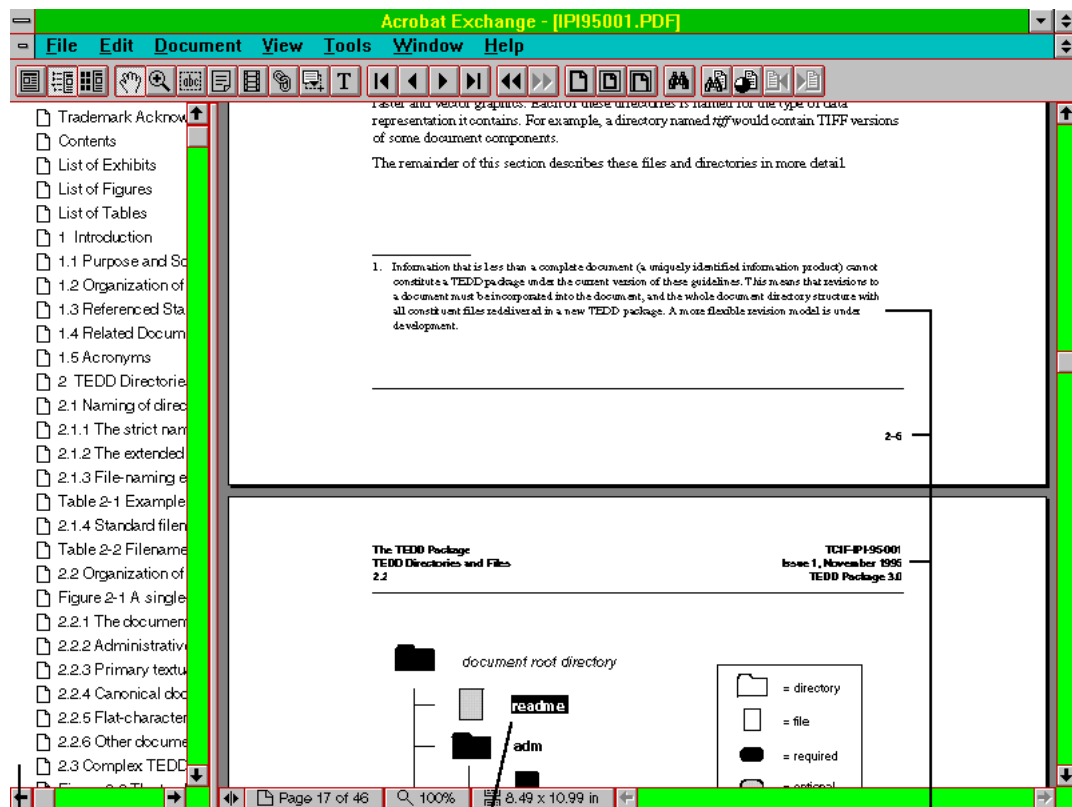
Figure 4-1 Sample Netscape (HTML) Screen
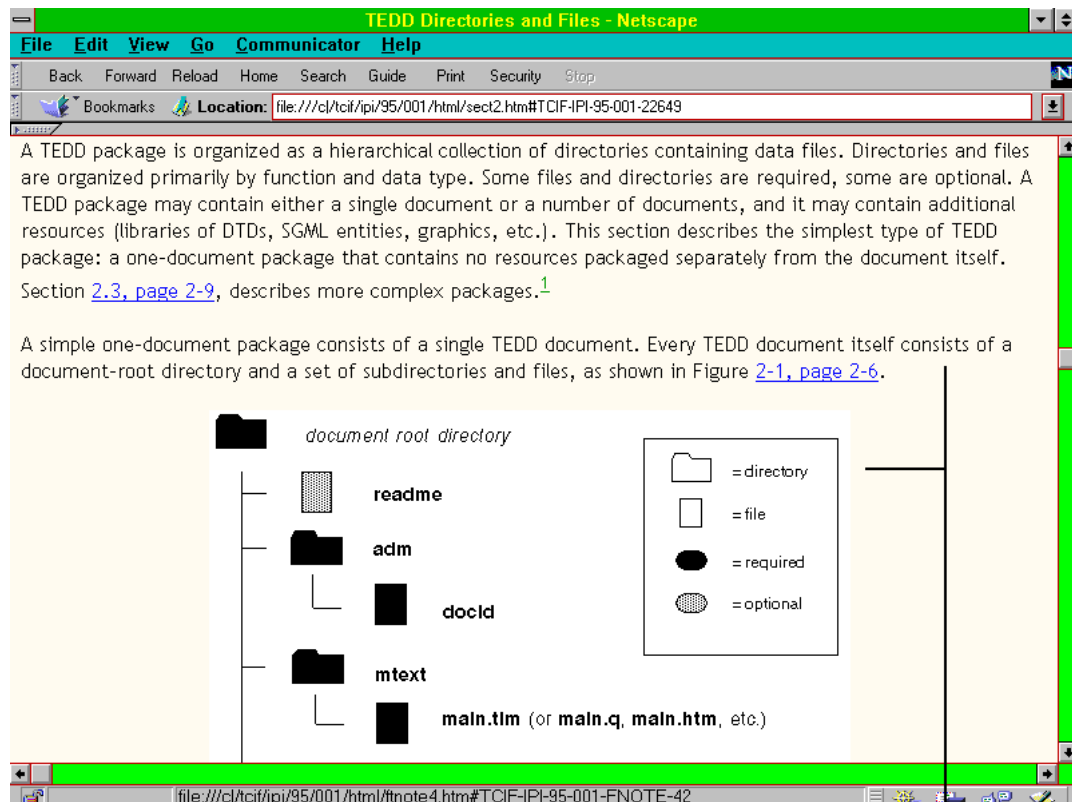
There is no table of contents unless created separately.

Text wraps to fit the window. Graphics are displayed pixel-for-pixel. The user can choose the basic font.

**Figure 4-2** Sample Acrobat (PDF) Screen

This "Navigator" (table of contents) can be created manually in Exchange or automatically with add-on tools.

Text search finds words in drawings.

The document looks like the original, including its headers, footers, margins, and type sizes. Enlarging to make the text easier to read may make lines too wide to fit the window. Hyperlink hotspots will be underlined or colored only if they are in the original.

Figure 4-3 Sample Panorama (SGML) Screen
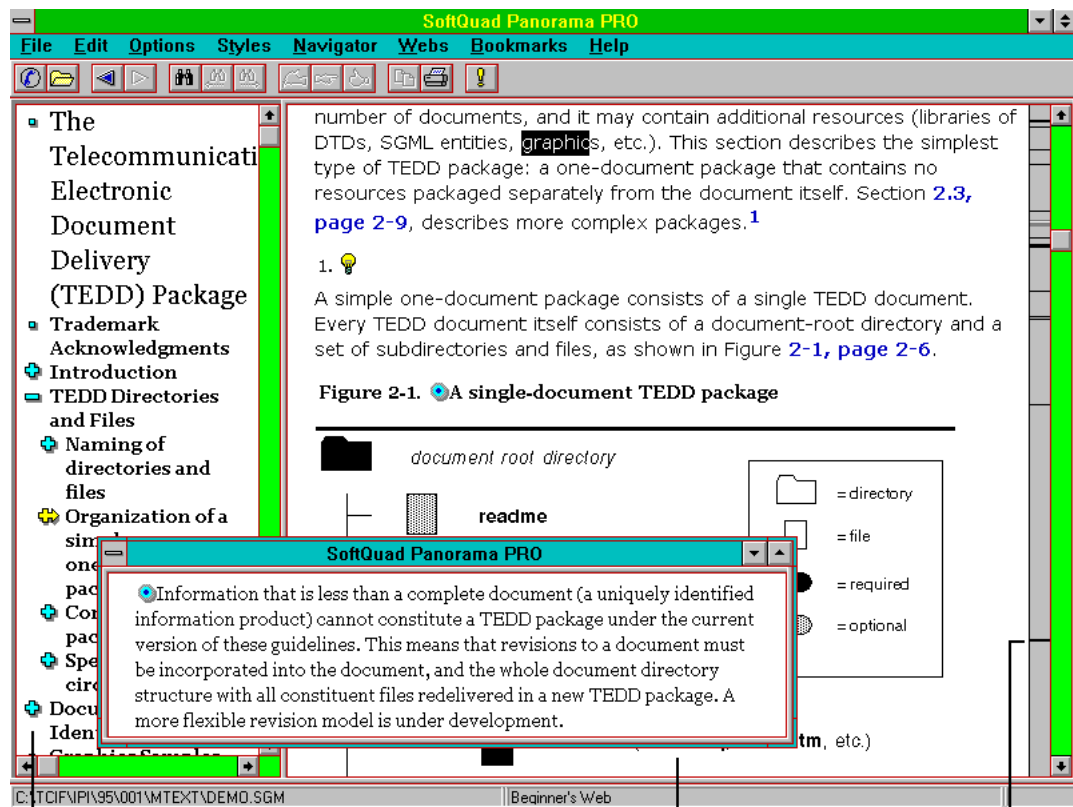
A telescoping table of contents (or a list of any type of element) can be created automatically.

Text wraps to fit the window. User-controlled stylesheets specify the font for each element. Any element (a footnote, in this case) can be made to appear in a pop-up window. Graphics are resized to the window, but can be "torn off" to separate windows.

The occurrence density display shows where a searched-for term occurs.