Produced for:

ASSESSMENT OF FILE FORMATS TO SUPPORT THE LIFE CYCLE MANAGEMENT OF ELECTRONIC PATENT AND TRADEMARK CASE FILES

February 26, 1999

Contract Number: 50-PAPT-700041 **Task Number**: 56-PAPT-8-05089

Deliverable: 98-03-08

Government Task Managers: Arthur F. Purcell, (703) 308-6868, FAX (703) 308-6916
Kathy Schultz (703) 308-7400, FAX (703)308-7407

Contractor: Cohasset Associates, Inc.
3806 Lake Point Tower
505 North Lake Shore Drive
Chicago, IL 60611
Tel. 312/527-1550

Contractor Task Manager: Richard D. Fisher (408) 741-1287, FAX (408) 867-1289
e-mail: rd.fisher@worldnet.att.net

FINAL

Table of Contents

1 INTRODUCTION

Between now and the year 2003, the Patent and Trademark Office (PTO) is undertaking an electronic patent and trademark application filing, processing and maintenance program that is part of the "Reinvention Goals for 2000" which establishes the PTO's vision for the 21st Century - - " to lead the world in providing customer-valued intellectual property rights that spark innovation, create consumer confidence and promote creativity."

The electronic patent and trademark business processes can be broadly divided into an applicant component, and a PTO case file processing and maintenance component. There is an underlying commonality of information and requirements that is shared between these two broad stages of the patent and trademarks records life cycle.

The degree to which standardized file formats are used to exploit this underlying commonality will have a profound effect on the customer acceptance and cost effectiveness of electronic patent and trademark application filing and processing activities. It is for this reason that this study on electronic file formats over the life cycle of electronic patent and trademark records is being performed.

Section 1 contains introductory and background information about the document. Section 2 defines the purpose and scope of the deliverable. Section 3 provides a summary of the approach, criteria and recommendations for management. Section 4 explains what file formats are, what they do, and why they are important to the PTO. Section 5 defines evaluation criteria that will be used in Section 6 to evaluate seventeen (17) file formats that have varying degrees of potential use in supporting the long-term management of the life cycle of electronic patent and trademark applications. A table that summarizes the evaluation of each standard concludes this section. Section 7 contains recommendations for specific file formats that the PTO should consider in the implementation application information systems (AISs) and the information technology infrastructure for managing electronic patent and trademark records.

2 PURPOSE AND SCOPE

While this report focuses primarily on assessing the file formats that are available for use in managing the life cycle of electronic patent records, the analysis, findings, and recommendations are intended to also be applicable to trademark records. This assessment is based upon evaluation criteria that take into account the requirements of the applicant and the PTO. For each of the file formats reviewed, general descriptive information is given followed by an assessment of the strengths and weaknesses, and a prognosis of its persistence over time. The report concludes with a recommended course of action for the PTO.

3 MANAGEMENT SUMMARY

This study assesses seventeen electronic record file formats relative to their ability to meet selected criteria for the long term accessibility of patent and trademark records over the full life cycle. File formats define the underlying content and structure of an electronic record in a manner that allows a computer to interpret and understand this information, thereby making the record processible and transferable.

- **Processible** refers to the ability, over the full retention life, to render an electronic record on a monitor or a printer, and includes the process of creating, storing and retrieving the record for that purpose.
- **Transferable** means the ability to move an electronic record across technology platforms while maintaining the integrity of content, structure, and context

The seventeen file formats reviewed are:

TEXT/COMPOUND FORMATS

- Standard Generalized Markup Language (SGML)
- Extensible Markup Language (XML)
- Portable Document Format (PDF)
- Rich Text Format (RTF)
- ISO/IEC 8211 – Data Descriptive File for Information Interchange
- ISO 8613 – Open Document Architecture (ODA)/Open Document Interchange Format (ODIF)

VECTOR GRAPHICS

- Scalable Vector Graphics (SVG)
- Initial Graphics Exchange Format (IGEF)
- Computer Graphics Metafile (CGM)

BIT MAP IMAGES - COMPRESSION

- Group 4 ITU (CCITT)
- Portable Network Graphics (PNG)
- Joint Bi-level Image Group (JBIG)
- Graphics Interchange Format (GIF)
- Joint Photographic Experts Group (JPEG)
- Lempel-Ziv-Welch (LZW)
- Motion Picture Experts Group (MPEG)

IMAGE FILE HEADER

- Tagged Image File Format (TIFF)

Criteria for assessing these file formats were developed by first laying out a logical process flow of patent application filing and patent processing and maintenance, then defining three categories for evaluation:

1 *User applications* – which covers the authoring, assembly and submittal of the application and the receipt of the granted patent (if granted).
2 *PTO applications* – which encompasses the different stages of the patent life cycle processing and maintenance and defines the requirements from the initial receipt and review of the application, through examination, publishing, maintenance and long term preservation.
3 *Technology Considerations* – covers areas such non-proprietary standards and the persistence of those standards.

Within these three categories, the general requirement areas used as criteria for performing the assessment are:

- **Preparation** – the ability of the applicant to author, assemble and submit a patent application and related electronic records in an cost/efficient manner using readily available, easy to use COTS software.
- **Navigability** – the ability to directly address any structural data element or locate any textual string in an electronic patent record.
- **Portability** – the ability to transfer an electronic patent record from one hardware and software environment to another while maintaining the integrity of the content, structure and context.
- **Multi-media** – the ability to support the internal structure of patent applications and related electronic patent records that may consist of text, vector graphics, bit map images or tables from a relational database, etc., either imbedded in the application or as links to external sources.
- **Integrity**- faithful rendering of the patent application as prepared and submitted by the applicant, as well as all submitted or generated patent case file records, with no loss of content or structure, for the full retention life of the patent case file, including any migration to other media or systems technology.
- **Rendering** – the ability to process the content and structure of all electronic records in the patent case file and faithfully reproduce the electronic record on a monitor or printer over the full retention life of the patent.
- **Format Persistence** – the ability of an industry or defacto standard to gain widespread user and vendor support, thereby offering a stable, sustainable, long term file format.
- **File Size** – the actual size of the file influences storage cost, transmission speed and rendering time.

Confidentiality, while an important legal requirement, cannot be assured as an inherent part of a file format, rather, other means such as encryption or access controls better serve this need.

Overall cost effectiveness of the file format is another consideration for both the applicant and the PTO. It is important that the applicant be able to prepare and submit electronic patent forms and related files, as well as retrieve and view published patent information using cost effective desktop applications or an Intra/Internet interface. It is also important for the PTO to receive, view, navigate, store, retrieve, publish and maintain electronic patent case files in a manner that provides the most reasonable cost to the PTO and, as such, the inventors and applicants. There are tradeoffs between the file formats reviewed in this study and, as such, there is no single file format that satisfies fully all of the criteria. On balance, XML and its extensions, including the newly proposed Scalable Vector Graphics (SVG) extension for graphic representations, appears to best satisfy the criteria except for one major problem, namely, the SVG extension is not likely to be adopted for over a year and currently there is very little XML software available. All of the evidence suggests that, as predicted by its proponents, XML and the SVG extension (as well as other XML extensions) will rapidly evolve to become the preferred file format for conducting electronic commerce. It is also assumed that XML will become as ubiquitous as HTML in terms of: creating XML output from desktop office productivity applications, sending/receiving e-mail in XML format, and being integrated (at no or low cost) as a viewing/navigation/reproduction capability into Internet browsers and desktop productivity applications. Based on the these overall considerations and the detailed assessment, the following recommendations are made:

- **Provisionally adopt Extensible Markup Language (XML) and its extensions, including Scalable Vector Graphics (SVG), for supporting electronic patent application preparation, review and examination.**

  The provisional nature of this recommendation is due to the current status of XML and its extensions, in that full scale implementation support has not been delivered and, as such, widespread user acceptance has not yet been established. It is expected that full vendor and user support will materialize over the next one to two years whereupon the provisional nature of this recommendation would be lifted.

  This recommendation is based on the ability of XML to support structural tagging and navigation and the ability of SVG, along with other XML extensions to provide support of graphics and more complex work units. Since XML can also recognize “well-formed documents”, non-tagged electronic records such as e-mail (that is not transmitted in XML) should also be able to be cognized and processed. XML and its associated extensions and capabilities (as currently defined or under definition) meets the majority of the assessment criteria, including those related to preparation, rendering, portability, multi-media, navigability and, presumable, format persistence. As such, XML should provide a single file format that can create, store and accurately render the content and structure of all electronic records contained in a patent case file. Potential drawbacks of this format may be the cost and ease of preparation for the applicant and the lack of any inherent preservation of integrity. In support of this recommendation, it is suggested that the PTO:

  – Develop an XML Document Type Definition that could be implemented and include the SVG extension so as to standardize the submission of electronic patents.

- Review the XCI project of the New Mexico Federal Courts and learn how XML with the SVG extension is being implemented.
- Develop and prototype software specifications to support the creation of XML with the SVG extension for electronic patent applications and related documents.

- **Provisionally adopt Extensible Markup Language (XML) and its extensions, including Scalable Vector Graphics (SVG), for electronic publication and dissemination of granted patents.**

  Since the as-filed and amended, if required, patent application and associated electronic records are recommended to be produced and stored in XML (with extensions) format, publishing of the patent in XML is the logical and cost-effective choice.

  One drawback with XML, as with SGML, is that the file format does not provide an inherent ability for ensuring the integrity of content and structure. If the requirement for providing inherent file integrity (non-revisable) of the published patent is deemed essential, and if a non-revisable method for rendering the published patent is not possible with XML (and the SVG extension), then PDF would be a better choice. PDF has the obvious disadvantage of being a proprietary standard and may be superceded over time by the .SVG extension of XML, particularly in Internet-based electronic commerce.

- **Provisionally adopt Extensible Markup Language (XML) and its extensions, including Scalable Vector Graphics (SVG), for long term maintenance and preservation of the electronic records in the patent case file for the full retention life.**

  Assuming that the XML and its extensions become the industry standard for electronic records produced in electronic commerce, the retaining the electronic patent case file records XML should meet the long term maintenance, use and preservation requirements related to format persistence, rendering and portability.

  Since XML does not currently provide an inherent ability for ensuring the integrity of content and structure, other measures must be taken by the PTO to meet this requirement.

- **Adopt Scalable Vector Graphics (SVG) as the standard for vector graphics.**
- **Employ only lossless compression methods for bit map images, including Group 4 ITU and JBIG for bi-tonal images, and PNG for grayscale and color images**
- **Define a "standard" TIFF image header (including any PTO-specific tags) and require the producers of electronic patent applications to use it when the submission includes bit map images**

4 FILE FORMATS AND WHAT THEY DO

4.1 FILE FORMATS DEFINED

In computer processing and storage all digital information, including multi-media electronic records -- text, vector data, still and moving images, spreadsheets, and databases-- are represented by a specified sequence of electrical representations of 1s and 0s, which is the language that computers understand. These 1s and 0s are called a bit stream. This bit stream also contains information about special operations (e.g., compression) that may have been performed on it, information that notifies the operating system how to interpret (e.g., text or bit map image) the bit stream, and software application specific instructions on how the record is to be rendered on a monitor or printer. This combination of information, instructions, and data is called a file format. A file format, therefore, refers to the underlying structure of electronic records that typically is not visible to the creators and users of the material. The ability of computers to interpret and understand file formats is crucial to ensuring the processibility and ability of digital information.

4.1.1 Processibility

Processibility refers to the creation, storage-retrieval, and rendering of electronic records on a monitor or printer. For example, a word processing software application such as Microsoft Office 97 inserts non-printing characters in the bit stream that denote margins, line breaks, type font, size of font, special characters, and the like. The software also adds a file extension (e.g. docs) that identifies what the bit stream represents. Each word processing software application has a slightly different file format that is not intuitively transparent to another word processing software application. The file format of records created in an Office 97 environment means that they can only be opened and rendered in an environment that supports Office 97 or in an environment that supports an automatic translation filter. In addition to domain specific application specific file formats for textual records, there are file formats that support the processibility of classes of other digital data, including vector graphics, bit map images, spreadsheets, and relational databases. Each one of these file formats supports the processibility of a specific class or family of material (e.g., textual or graphics) in a domain-specific software application.

Chemical formulas dealing with the composition of matter and biological sequences represent more narrowly defined classes that tend to be either industry or discipline specific. Specific file formats have been designed that support the processibility of a number of chemical expressions and biological sequences. Other PTO studies have already addressed these file formats for chemical formulae and biological sequences (Draft Complex Work Unit (CWU), June 4, 1998).

4.1.2 Interchange

The purpose of an interchange file format is to provide a content-independent and media-independent vehicle to move digital information across different technology platforms without any loss in structure, content, or context.. Data interchange formats can be characterized as general purpose, specific, and transparent. Some general purpose data interchange file formats, which may require a special purpose program, support the transfer of content or media dependent records to a neutral (i.e. application software and hardware independent) format where they can be held and then moved to any target application that can recognize and interpret the data interchange file format. Other software applications provide a transparent export or import function that can be invoked with one click on a screen icon. Examples of the latter are Lotus Data Interchange Format (DIF) and Microsoft SYmbolic Logic Link (SYLK) for spreadsheets. Some data interchange formats are automatic in the sense that no or little user involvement is required because the export or import software functionally recognizes the source or target software and takes the appropriate action..

5 EVALUATION CRITERIA FOR FILE FORMATS

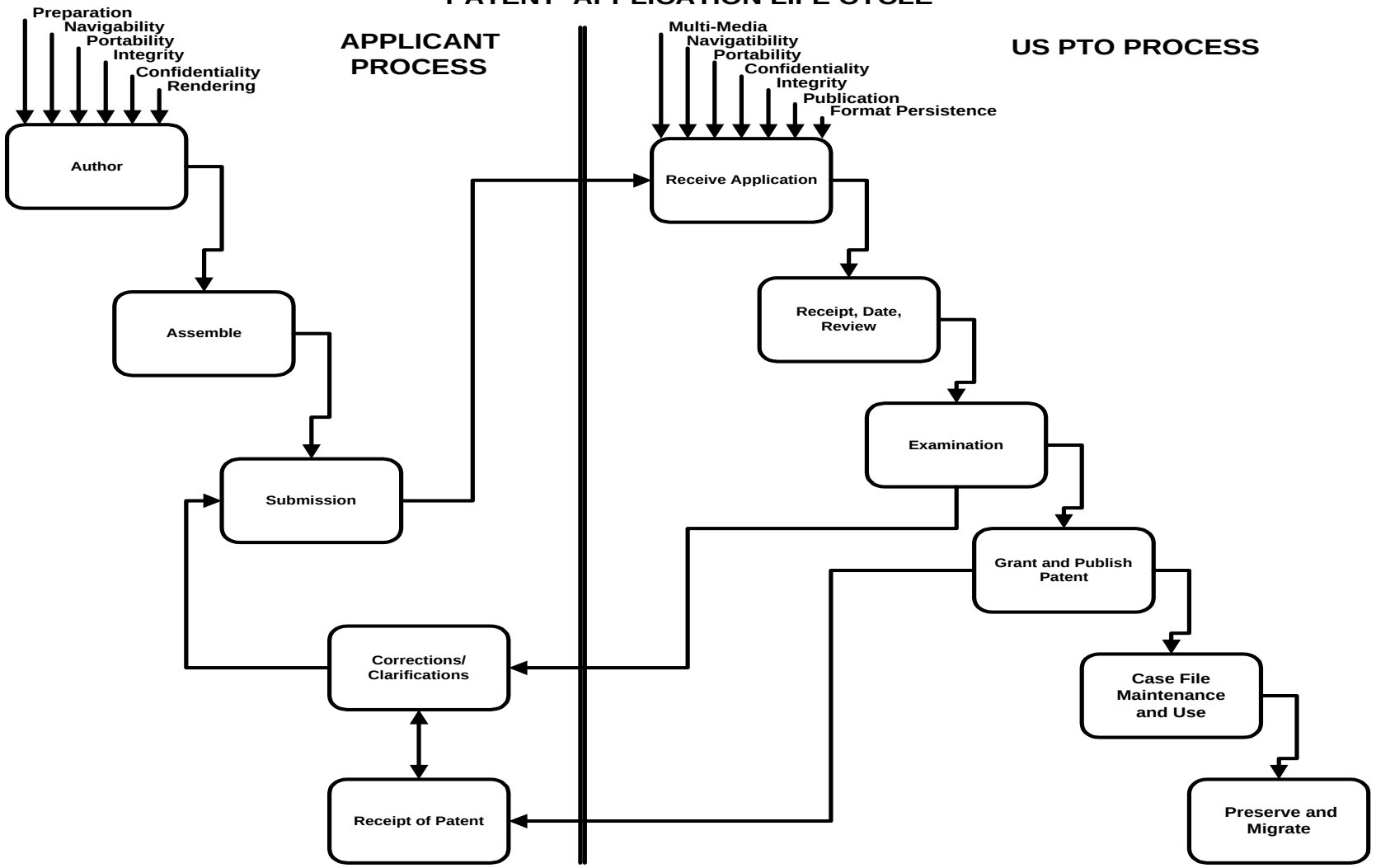
This section of the study identifies a number of criteria for evaluating electronic file formats that support the patent case file life cycle. These criteria are reviewed in three different contexts: applicant considerations, PTO considerations, and technology considerations. The applicant and PTO areas are first depicted in the context of a logical process model of the electronic patent application life cycle and then are elaborated in some detail relative to specific evaluation criteria

5.1 LOGICAL PROCESS MODEL

The examination of electronic patent applications is a very complicated process and it is easy to get bogged down in detail that is not essential to understanding the process. The life cycle of an electronic patent application, which was previously identified (Legal and Records Management issues and Requirements Related to the Electronic Filing of patent Applications, November 21, 1997) focuses attention at a higher conceptual level. A logical process model also can be helpful because its purpose is to increase the understanding of complex phenomenon by eliminating any detail that does not affect its behavior, such as who does what and how it is done. A logical process model of the electronic patent application life cycle has been created in order to identify clearly the inputs, requirements, and outputs associated with the examination and approval/disapproval, or abandonment of an electronic patent application. This process model is shown in Figure 1.

The process model is divided into two components: the applicant process and the PTO process. Each of the boxes denotes an activity that is carried out. Just above the Author box and Receive Application box there are vertical arrows pointing down. These arrows denote requirements that cascade down through the processes. In the Applicant process there are six vertical arrows pointing downward: Preparation, Navigability, Portability, Integrity, Confidentiality and Rendering. In the PTO process there are seven vertical arrows pointing downward: Multi-Media, Navigability, Portability, Confidentiality, Integrity, Publication, and Format Persistence. Each of these requirements is important but format persistence is especially important because of its impact on Case File Maintenance and Use, and Preservation and Migration over the full retention life of electronic patent case files.

FIGURE 1
LOGICAL PROCESS MODEL OF THE
PATENT APPLICATION LIFE CYCLE



5.2 APPLICANT PROCESS

There are five stages in the life cycle of the applicant process. They include author, assemble, submission, and corrections/clarifications. The fifth stage, which is not shown n the process model, involves applicant rendering of the electronically published patent grant. Consequently, customer applications criteria are divided into three categories (there are six “down arrows” on the chart related to the application process).

5.2.1 Author/Assemble

The first two stages of the applicant process involve authoring the various components and then assembling them into a final product that is ready for submission to the PTO.

5.2.1.1 Preparation

Preparation should be done with readily available, easy to use software that seamlessly integrates all of the file formats that are appropriate for the specific patent application being prepared. In this context seamless integration means that a particular file format can be opened within an application, say word processing, without exiting to a program that launches software that implements the file format. Typically, this would entail clicking on the appropriate screen icon.

5.2.1.2 Navigability

It is likely that the person preparing an electronic patent application may need a navigation tool that supports direct access to any part of the application, including the external linkages discussed above. The preparation software used to author and assemble an electronic patent application should permit direct navigation of structural (tagged) elements and full text (Boolean) navigation of the patent application and other records in the case file.

5.2.2 Submission

When the preparer of an electronic patent application is satisfied that it is complete, it is submitted to the PTO for review and examination. Submission, of course, means electronic transfer, which involves three critical areas for each patent record: portability, integrity, and confidentiality.

5.2.2.1 Portability

Portability means that patent applications and related electronic records created in one software and hardware environment can be transferred to a different software and hardware environment without any diminution in processibility or loss of content and structure. For the applicant this means that a patent application produced in one software and hardware environment can be submitted to the PTO, confident that the patent application can be "opened" and processed by PTO computers and software. Portability of electronic patent applications can be supported by the software used to author and assemble them or by a specific interchange format.

5.2.2.2 Integrity

Closely related to electronic patent application portability is integrity, which in this context means that the exact view or a faithful representation of the view the creator of the patent application had at the time of submission is the same view that the PTO receives and processes. Specifically, this means that the electronic patent application must be in a file format that permits its faithful reproduction on any operating system or computer architecture supported by the applicant and the PTO. As such, the file format used must retain all of the original formatting such as graphics, tables, type fonts, and other specialized symbol representations.

In addition, there is always the potential for errors to occur in the transmission of electronic records that can undermine integrity, however, cyclical redundancy checksums and hash digests can be used to detect when such errors occur. The greater risk in record integrity is likely to occur when bit map images are part of an electronic patent application. For example, if a lossy compression algorithm (i.e., JPEG) is used to store a bit map image this means that the bit map the PTO receives will not be an exact replication of the original bits. This loss of detail resulting from the use of a lossy compression algorithm can never be restored. Of course, the argument can be made that this irreversible loss of detail may not be easily detected by human vision and therefore is of little consequence.

It should be noted that file formats as such can not protect or ensure the integrity of electronic patent applications. In many instances, a cyclical redundancy checksum or a hash digest of an electronic patent application can be used to determine if any alterations have occurred during transmission or since the last time the there was any activity associated with the application. Also, some advocates of electronic records integrity support the use of Portable Document Format (PDF) because it is "unrevisable." As will be noted later in this report, this is more apparent than real. More to the point, however, is that rigorous quality control procedures and audits are likely to provide a much stronger basis for ensuring electronic patent application integrity.

5.2.2.3 Confidentiality

A primary concern for many individuals who submit electronic patent applications will be the protection of the confidentiality of their application. It is one thing to submit a patent application by certified mail and quite another to submit an electronic patent application via the Internet because open networks, such as the Internet, are considered to be vulnerable to intrusion unless a powerful tool, such as encryption is used.

However, no file format in and of itself can ensure the protection of the confidentiality of a patent application. There are several widely available tools and techniques such as encryption and one-way hash digests that support confidentiality but they are not considered file formats as such. Confidentiality also can be controlled through implementation of rigorous quality control procedures, audit trails, and compliance audits.

5.2.3 Receipt of Grant

Once a patent has been granted, it is published. Part of this publication process involves delivery of a copy of the approved patent grant to the applicant for review. Implicit in this process are three key issues: the portability, integrity, and faithful rendering of the approved patent grant.

5.2.3.1 Rendering

Presumably, publication will consist of sending a copy of the approved patent grant to the applicant and posting of the published patent grant on the PTO World Wide Web site. From the applicant's perspective it is extremely important to be able to faithfully reproduce the grant on a monitor or printer with no loss of content or structure. The wide diversity of technology platforms and software applications in use today and for the foreseeable future means that the rendering of approved patent grants must be hardware and software independent.

5.2.3.2 Portability

As noted earlier, portability means that electronic patent applications created in one software and hardware environment can be transferred to a different software and hardware environment without any diminution in processibility or loss of content or structure. For electronic patent applications, as well as any other submitted patent records, to be portable, they should be fully processible and capable of being rendered for viewing or printing with any widely used and supported operating system and application software package. The file format(s) used for electronic publication must support this level of portability. In addition, the software that supports portability should be widely available and easy to use.

5.2.3.3 Integrity

Under current procedures the preparation of a granted patent application for publication is outsourced to a service bureau that uses a Standard General Markup Language (SGML) template to standardize its physical representation, that is, how the published patent application looks. The file format used for electronic publication must ensure that when a published patent is returned to an applicant it is a faithful reproduction of the patent application that the examiner approved and then was prepared for publication. This same level of integrity is necessary for those interested individuals who retrieve a published grant from the PTO WWW site.

5.3 PTO APPLICATION CONSIDERATIONS

In the patent application life cycle model there are five stages: receipt and review, patent examination, patent grant and publication, case file maintenance and use, and preservation and migration. The primary areas and criteria for the PTO process, as represented by the seven vertical arrows pointing downward in the logical process model are:

5.3.1 Receipt and Review

In the receipt and review stage, patent applications (and related electronic records) are received, dated, and reviewed for completeness. Completeness means that the required elements of the patent application have been submitted by the applicant and that other requirements, such as receipt of the filing fee, are satisfied in a timely manner.

5.3.1.1 Multi-Media

The PTO should be prepared to deal with multi-media electronic patent applications that may consist of text, vector graphics, bit map images, or tables from a relational database. In multi-media or compound electronic document applications these components (bit map images, vector graphics, and tables from a database) may be directly embedded in the application or links or pointers to external sources that are transparent to users are embedded in the text. Structural components such as bit map images embedded in application would be automatically rendered when the document is opened while structural components that have pointers to an external source could be rendered only when the application is available. Given this multimedia dimension of electronic patent applications, the file formats that support the internal structure of assembled patent applications must not preclude the use of linkages or pointers to components that reside in an external application.

5.3.1.2 Confidentiality

An underlying requirement of all but the preservation and migration and case file maintenance and use stages is the need to protect the confidentiality of the patent application. As noted earlier in the Applicant process, no file format in and of itself can ensure the protection of the confidentiality of a patent application. There are several widely available tools and techniques such as encryption and one-way hash digests that support confidentiality but they are not considered file formats as such. Confidentiality also can be controlled through implementation of rigorous quality control procedures, audit trails, and compliance audits.

5.3.1.3 Portability

From the PTO perspective, portability of electronic patent applications means that all patent applications received must be capable of being automatically opened and placed in the workflow process without any loss of information content or structure. The file format used in the creation and transfer of an electronic patent application, therefore, must be one that both meets the requirements of the patent examination work flow applications and supports portability so that the electronic patent applications are application and domain technology independent.

5.3.1.4 Navigability

PTO officials anticipate that the intake processing and examination processes will be automated or automation assisted to the highest degree possible. This will require an electronic template for patent applications that is composed of a series of required pieces of information that conform to certain PTO defined specifications. Each one of these required pieces of information must be directly addressable in the form of structural data elements that can be automatically navigated and, where appropriate, automatically processed. In effect, automatic navigation could consist of comparing a check list of required information in a specified format with the actual structural content of an electronic patent application. The file format used in the preparation and transfer stages of a patent application, therefore, must not preclude or otherwise impede this automatic navigation.

5.3.2 Examination

The core of the PTO process of determining if a patent application can be granted occurs in the examination process.

5.3.2.1 Portability

Portability in the examination process means that electronic patent applications are media and technology platform independent but with an added emphasis on being able to deal with multimedia or compound electronic patent applications. Patent examiners will need to open and render for display or printing all the components of a multimedia electronic patent application (e.g., bit map images). There are three ways this can be done. One way is to have access to the version of the application(s) used to create the components. A second way is to have viewers that can automatically open the components. The third way is through the use of software that can import the specific components into the PTO application environment.

5.3.2.2 Navigability

In the examination stage there is no uniform set of procedures or steps that each patent examiner must follow. Consequently, patent examiners are likely to develop idiosyncratic examination work sequences. In order to accommodate such variations in examination work sequences, flexible navigation of the structural elements of a patent application is a base line requirement. Flexible navigation means support for Boolean searches at the structural level and full-text retrieval. This translates into the criterion that file formats must not preclude or impede Boolean searches and full-text retrieval.

5.3.3 Grant and Publication

If, during the examination step, it is determined that the patent application meets the criteria for allowance, the applicant is notified of the grant, and the patent is issued. As described earlier, the current publication practice involves outsourcing to a thirty party for conversion of the approved patent to a PTO defined Standard General Markup Language (SGML) template that standardizes its physical representation, that is, how the published (printed) patent application looks. When the electronic patent application process is implemented, PTO officials anticipate

being able to directly publish the patent from the electronic records of the as-filed patent and any amendments made during the course of examination.

5.3.3.1 Integrity

Protecting the integrity of electronically published patent grants is a fundamental requirement that means they should have the same logical and physical representation, that is, in how they look when they are rendered on a screen or a printer regardless of the operating system or technology platform is a fundamental requirement. Equally as important is the requirement to protect electronically published patent grants from accidental or intentional alteration.

5.3.4 Case File Maintenance and Use

This stage of the life cycle of electronic patent applications begins at the time of the issuance of a patent and extends for the full retention life of the patent. Granted patents case files are maintained permanently, with responsibility for maintenance being transferred to the National Archives and Records Administration (NARA) years after the date of filing. Abandoned patent application case files are maintained for 20 to 23 years, depending on the date of filing.

5.3.4.1 Examination

During the examination stage of electronically filed patents, there is no single prescribed way for an examiner to proceed. Consequently, examinations are likely to be ad hoc with little predictably of what examiners will want to see or exactly what the query and retrieval sequence will be. Nonetheless, it is reasonable to conclude that the file formats must support a search and retrieval capability that provides structural access (i.e., a structural component such as date or type of patent) and full-text (Boolean) access to electronic patent application records.

5.3.5 Preservation and Migration

All electronic records in patent case files must be retained in processible form for the full retention life. During this time, it is certain that technologies will change and the PTO will implement successive new technologies. Ensuring long-term access to electronic patent case files as technologies change will require the PTO to migrate the electronic records in patent case files to the current technology environment. Repeated migration of electronic records in which their underlying bit streams could undergo change raises a question of how to ensure that there is no loss in information content or structure. Deliverable 98-03-10, *Long-term Access and Migration Strategy for the Life Cycle Management of Electronic Patent and Trademark Case Files*, addresses this issue.

5.4 TECHNOLOGY CONSIDERATIONS

There are three evaluation criteria that fall under the heading of technology considerations: file storage requirements, standards (non-proprietary/industry de facto) for file formats and persistence of file format standards.

5.4.1 Non-Proprietary Standards

The prevailing model of computer architecture today is open systems, which denotes a multi-vendor product environment in which users can choose between a variety of compatible products that perform the same tasks and support the interchange of digital information. This multi-vendor product environment is based upon non-proprietary information technology standards that make it possible for heterogeneous computers and computer peripherals to communicate and process digital information with no loss in content or structure.

Generally speaking, there are two kinds of non-proprietary information technology standards. The first consist of standards that are the product of an authorized standards body such as the American National Standards Institute (ANSI) and the International Standards Organization (ISO). Typically, authorized standards have a broad base of user and vendor community support. The second kind of non-proprietary information technology is called an industry de facto standard, which means that a dominant player in a specific industry has developed a technique or tool that is so widely implemented it has the force of a standard but without the official authorization of a standards setting body. One example is the Rich Text Format developed by Microsoft to support the interchange of word processing documents.

The first criterion for evaluating file formats is if they are non-proprietary. In all but the most special circumstances, such as file formats for chemical expressions, a proprietary file format should not be used.

5.4.2 Format Persistence

The second evaluation criterion under the heading of Technology Considerations is the projected persistence over time of a non-proprietary/industry de facto standard. The greater the number of software implementations that provide users with choices the greater is the likelihood of widespread usage, which is essential for the persistence of a standard. Absent this level of software implementation, even well thought out and carefully designed standards will have limited utility. Therefore, the best single predictors of file format persistence are multi-vendor software implementations and user satisfaction with the software implementations.

Perhaps the most persistent information technology standard is the American Standards National Information Interchange code because there is virtually no computer today that does not support ASCII. On the other hand, consider ISO 8613, "Office Document Architecture/Office Document Interchange Format (ODA/ODIF), which was approved in 1988. The purpose of ODA/ODIF (ISO 8613) was to facilitate the exchange of office documents such as reports, letters, memoranda, and the like between dissimilar computer platforms and applications. ODA is a robust standard that defines a business document architecture in terms of content and two hierarchical structures: a logical structure and a layout structure. In addition, ODA supports character data, raster, and picture data. ODIF defines the encoding of a bit stream that can be moved from one system to another. However, there were only a couple of pilot software implementations in Canada and Europe and the standard never penetrated the market place. One reason why it did not is the rapid and somewhat unexpected wide spread use of SGML. What happened to ISO 8613 clearly demonstrates the consequences of limited market acceptance.

5.4.3 File Size

In some instances the use of a specific file format can have a direct influence on the size (number of bytes) of an electronic patent case file. In turn, this can affect storage costs, transmission speed, and rendering time. For example, moving a textual file in Word 97 to the Rich Text Format (reviewed later) typically results in doubling the number of bytes in the file. Clearly, the magnitude of this increase in file size could have a significant impact on the information technology infrastructure of the PTO.

6 EVALUATION OF FILE FORMATS

Earlier attention was called to multi-media aspects of electronic patent applications or compound electronic patent applications and a distinction was made between structural components that are embedded in an electronic patent application and pointers or links to external sources that are embedded in an electronic patent application. Typically, the former occurs when the software package used to create and assemble an electronic patent application supports an appropriate rendering capability as in the case of embedding vector drawings, such as VISIO, in a Word document. External links or pointers to a vector drawing or bit map image, for example, would require access to the application (e.g., VISIO) in which they were created or maintained. For the purposes of this study, therefore, text should be understood as alphanumeric data that may contain embedded objects and/or links and pointers to external objects.

6.1 TEXT/COMPOUND FORMATS

6.1.1 Standard Generalized Markup Language (SGML)

From the beginning of the PC revolution, one of the mostly widely used applications has been word processing systems to create, edit, print, and store textual documents. Word processing packages such as WordStar, MacWrite, WordPerfect, and Microsoft Word allowed users to compose text and to select different type fonts and sizes, spacing, margins, and indenting, which were then embedded in ASCII text as non-printable formatting instructions. Each word processing software application tended to handle these formatting instructions differently so it was not possible, for example, to use Microsoft Word to read a document created in WordStar. In the early 1980s the publishing industry developed a standard description of electronic text to take advantage of text processing technology. This description separated physical representation from logical structure. About the same time, IBM was developing a Markup Language to standardize its in-house publication program so that all IBM publications would look alike. As a result of these two projects, a number of people began to realize that traditional markup of text tended to match the elements of its logical structure. If a standard way to separate the logical structure of documents from formatting instructions associated with these documents could be developed, it would have several important benefits. One benefit is that documents created in this “standard way” would be software and hardware independent and could be processed on any modern computing platform. Another benefit is that “standard electronic documents” or portions of them could be reused without any additional markup expense.

Interest in a standard spurred work on an international standard. In 1986 the Standard Generalized Markup Language (SGML) was adopted as ISO 8879. SGML uses a standard character set -- seven bit ASCII -- so a SGML document can be read and processed by any computer system. SGML consists of rules that define the logical elements or structure of a group of similarly structured documents and rules that identify these elements. Users of these documents can then describe the logical elements in what is called a Document Type Definition (DTD).

SGML does not mandate any specific DTD. Rather, it prescribes the rules and language whereby users can define a DTD for documents sharing a common structure, such as letter or a memorandum. The logical structure of a document consists of elements that are enclosed within brackets. A typical memorandum, for example, would consist of a date, a to (addressee), a from (the author of the memo), the subject of the memo, the main text, and a signature. The DTD for such a memorandum would identify the document type as a memorandum and would contain the logical elements of date, to, from, subject, text (paragraphs), and signature.

A DTD can also include "anchors" for internal and external hypertext links. Internal links permit direct and immediate navigation to user specified (i.e., the creator) parts of a document. External links can point to bit map images, vector drawings, spreadsheets, and databases, among others, that are attachments to the base document (text) or to an external source, such as a file server. Through the use of external hypertext links SGML can support multimedia or compound documents. In addition, the underlying ASCII text of SGML encoded documents can be used with any standard search and retrieval engine to locate single words or multiple words in one or more documents.

The actual physical representation, that is what a document looks like, is separated from the logical markup language. For this purpose, a style sheet is created that defines how a specific class or type of document will be formatted for rendering on a display monitor or printer. Sometimes this style sheet is called a Formatted Output Specification Instance (FOSI).

SGML is a mainstream technology platform. It is used widely in the publication of books, reports, and manuals. Through the Continuous Acquisition and Life - cycle Support System (CALS), the United States Department of Defense mandates the use of SGML in the acquisition, production, and distribution of technical manuals and reports, most of which are delivered under government contracts. SGML is one of the text formats approved by NARA for archival preservation. The Text Encoding Initiative (TEI), which is a major academic initiative to convert historical sources into computer processible form, uses SGML. SGML can be used in the preparation of office documents, although to this point the effort is modest. However, some records managers and archivists are advocating the use of SGML in the creation of office documents and records because it is virtually technology independent.

Because SGML documents are in a standard file format, it will not be necessary to convert them when hardware or software becomes obsolete as long as SGML software is operational. SGML embedded documents can carry all of the information necessary to use, display, or print them. SGML should be of particular interest to the PTO because of its potential for extending the longevity of electronic patent applications through protection from technology obsolescence and its ease of navigability with the appropriate search engine.

There are numerous software packages available that conform to SGML and permit users to define DTDs and to convert word processing documents to SGML documents. One such software package is SoftQuad, which is relatively inexpensive. Corel Word Perfect 8.0 also has a feature that converts word processing documents to SGML. However, it is necessary to create the functional equivalent of a Document Type Definition (DTD) and a style sheet for the physical layout of a specific type of document.

SGML does not mandate any specific DTD so users who wish to share or exchange documents must use the same DTD. If the PTO were to adopt SGML as the standard file format for submission of electronic patent applications, then a document type definition and style sheet

would have to be developed that patent applicants would be required to use when creating an electronic patent application.
So far as file size is concerned, the use of SGML probably would not result in a significant increase, although it must be acknowledged that this would depend largely on the complexity of the Document Type Definition (DTD). As a general rule, the use of a DTD and embedding SGML tags in the text could result in an increase in file size on the order of 3 to 5 percent.

STRENGTHS

- National and international standard
- Highly flexible and adaptable to PTO requirements
- Logical layout of documents would support structural navigation
- ASCII text could be searched with standard query and retrieval engine
- Retains processibility so it would be fairly straightforward to edit or revise SGML encoded electronic patent applications
- A small and virtually inconsequential increase in file size
- Easy to convert SGML encoded electronic patent applications to new information technology platforms
- Software implementations are widely available, particularly in Corel WordPerfect 8.0
- Wide market place penetration indicates substantial persistence over time
- Extensible with XML

WEAKNESSES

- Revisable text with no internal evidence of change that can raise questions about the integrity of electronic patent applications
- Requires a specific Document Type Definition and Style Sheet
- Conversion of a word processing document to SGML is not transparent
- Not widely used in the creation of electronic textual records and documents

6.1.2 Extensible Markup Language (XML)

On February 10, 1998 the World Wide Web Consortium (W3C) approved version 1.0 specification for XML. XML is a non-proprietary standard that is intended to make it easier to create, manage, and distribute SGML defined documents on the World Wide Web. Although it omits some of the more complex and less-used features of SGML that made it difficult for inexperienced users to implement, it nonetheless is still SGML. It employs the begin and end codes (< >) to denote tags that define components of the logical structure. These tags add intelligence to electronic documents because XML software will be able to distinguish between instances of the same character string as in the case of <author> Paul McCartney</Author> and <Beattles>Paul McCartney</Beatles>.
XML retains the Document Type Definition (DTD) of SGML and at the same time introduces a new class of documents called 'well-formed" or "standalone" that do not require a pre-defined DTD. A "well-formed" XML document is one that has a simple and unambiguous hierarchical structure that an XML "reader" can interpret. For example, the memo DTD for an SGML encoded memo described earlier is not necessary for a XML encoded memo. Instead, the tags

are inserted at the appropriate places in the text of the memo and the XML reader "infers" the logical structure. XML also has a NOTATION and ENTITY mechanism that supports graphics. It does not exclude any particular graphics file format, although it is likely that software implementations will at minimum support GIF, JPEG, TIFF, and CGM. Another feature of XML is that "domain-specific vocabularies" can be developed that any XML reader (i.e., parser) can understand. A number of domain-specific vocabularies, which sometimes are called XML extensions, are currently under development. They include a mathematical, chemical, musical, an astronomical, and genealogical markup language, among many others.

Complex XML documents require a style sheet for rendering, just as SGML documents do. A new Extensible Style Language (XSL) is being developed specifically for use with XML that incorporates formatting features from DSSL, the Document Style and Semantics Specification Language (ISO 10179) that is used with SGML.

Although there are many pilot or demonstration software projects that implement XML, there is very little fully implemented XML software currently available. The developers and promoters of XML believe that the benefits of using XML for World Wide Web publishing are so great that it is only a matter of time before multiple vendor sources for XML based software, including freeware, are widely available. As XML tools become widely available, XML and its family of extensions are likely to become a major tool in electronic commerce.

As multiple software implementations of XML become available and new extensions are added, it is likely that XML will displace SGML as the "lingua franca" of electronic commerce. The general assessment of most knowledgeable people is that by the year 2000 XML will be firmly established as the primary means for storing and transmitting portable electronic documents. It is unclear at this point whether the DTD that the PTO now uses for publication of approved grants is too complex for use with XML. This is an area that requires further investigation. Nonetheless, it does appear at this point that XML should receive the serious consideration of the PTO.

It is unlikely that the use of XML in the life cycle management of electronic patent applications and case files would have any adverse affects. Earlier it was estimated that use of SGML would result in an increase of 3 to 5 percent in file size. However, XML is less complex than SGML and in some instances ("well-formed documents") do not require a DTD so any increase in file size would be nominal at best.

STRENGTHS

- A non-proprietary W3C file format standard
- Relatively simple to understand and use
- Navigation of electronic patent applications based upon structural components
- ASCII text could be searched with standard query and retrieval engine
- Many software applications, including freeware, are likely to be available
- XML encoded documents would retain processibility so it would be fairly straightforward to edit or revise them
- It should be relatively easy to convert XML encoded electronic patent applications to new information technology platforms
- A small and virtually inconsequential increase in file size

WEAKNESSES

- Revisable text with no internal evidence of change that can raise questions about the integrity of electronic patent applications

- Requires a specific Style Sheet for rendering electronic patent applications
- Transparent conversion of a word processing document to XML is not currently available
- Only pilot demonstration software implements are currently available but this is expected to improve rapidly
- Not yet widely used in the creation of electronic textual records and documents

6.1.3 Scalable Vector Graphics (SVG)

Scalable Vector Graphics, even though primarily focused on graphic information, is covered in the Text/Compound Format areas, and proceeding the XML assessment, for two reasons:

- SVG is an extension to XML and may provide a more integrated perspective if addressed subsequent to the XML assessment.
- SVG provides for addressing text within the context of a graphics element and, as such, provides a limited “compound” format capability.

In May of 1998 Adobe proposed a Precision Graphics Markup Language (PGML) to the W3C for consideration as an extension of XML. Although PGML employs the two-dimensional model of Postscript and PDF it would be written in XML and therefore would not be a proprietary standard. A few weeks later Microsoft, along with Hewlett-Packard and Visio Corporation, submitted another vector graphics format called Vector Markup Language (VML0. The W3C technical committee established to work on this proposal decided to avoid any proprietary product as such. It has proposed an XML-based standard called Scaleable Vector Graphics (SVG) that incorporates features of PGML and VML.

SVG will employ the concept of x and y coordinates for locating the beginning and end of lines, shapes, and the like. Although SVG is intended to support vector graphics, it nonetheless can be used with associated textual information. Text can be embedded at any "name space" (i.e., an identifiable structural component such as the name of a circle or shape) which means that structural components of vector graphics can be easily navigated. As such it incorporates the faithful reproductive capability of PDF with a navigation capability like XML. The project director for the New Mexico federal courts pilot PDF/EDI electronic case file submissions stated that the project will incorporate PGML (or its equivalent SVG) into the project. With XML and SVG, the electronic application filing and processing and overall case file electronic records management could incorporate the best features of PDF and XML and make it possible for vector graphics files to be fully portable across different information technology platforms. According to Jon Ferraiolo, a senior computer scientist at Adobe who is chairman of the Technical Committee writing the SVG specification, formal approval by W3C is not likely to occur until December 1999. Even if there is some slippage in the scheduled completion of SVG this should have little effect on the deployment of electronic patent systems by the year 2003.

STRENGTHS

- A multi-purpose file format that is intended to support vector and text drawings
- When approved it will be a non-proprietary W3C file format
- Utilizes XML, which already is a W3C approved, non-proprietary standard
- Supports hypertext links and other navigational tools
- Compatible with PDF encoded documents

- SVG encoded documents are processible and thereby are easily transferred to new technology platforms over time
- Potential for major market presentation

WEAKNESSES

- The SVG file format standard will most likely not be submitted to W3C until late 1999.
- There are no software implementations or demonstration prototypes of SVG and consequently the cost for such software is yet to be determined
- Although SVG shows considerable potential and would meet most of the criteria for electronic patent application processing and case file management, it could not be realized for several more years

6.1.4 Adobe Portable Document Format (PDF)

With the exception of SGML, XML, and HTML documents, all the currently available software packages for spreadsheets and word processing embed instructions within documents on how they are to be rendered on a screen or a printer. Although these instructions are in the background and users may not be aware of them, in point of fact to display or print a document the original software used to create the document, a backward compatible version of the software, a run-time version of the software or a generic viewer that accurately interprets the embedded instructions must be used. This software dependency works against records portability.

Adobe Systems has developed a proprietary suite of software tools called “Acrobat” that allows users to open, view, browse, and print records as they appeared to their creators or recipients. Acrobat operates independently of the original software application, hardware, and operating system used to create electronic materials. Documents and records produced via Acrobat, called Portable Document Format (PDF) files, can contain any combination of text, graphics, and images in a device-independent and resolution independent format. PDF documents can be one page or hundreds of pages. They can be very simple or very complex. No matter what software package is used to create documents, they can be easily converted to PDF using Acrobat.

Within Adobe Acrobat there are two software tools available for creating PDF documents. The first involves converting text documents created in a word processing system such as Microsoft Word 97 to PDF by opening the 'Print" option and selecting PDF Writer. There are various options within PDF Writer, such as resolution and what is called "down sampling" that affect the size of a PDF file. In most instances, a PDF file created by PDF Writer will always be smaller than the original. The second tool for converting documents to PDF involves opening a “Distiller” that converts documents into Postscript, a device and software independent language for displaying or printing documents. Because Postscript is a language used for displaying or printing documents, detailed instructions about type fonts are embedded in the document. In addition, Postscript is based upon a two-dimensional data model in which x and y coordinates are used to specify the exact location on a page where a line or text begins and ends. These x and y

coordinates are also embedded in the document along with any vector data or text. Acrobat then takes the Postscript file (images, graphics, tables, and text that is bolded, underlined, or italicized with multiple font sizes and appearances) and converts it to a specially coded file called PDF. The Catalog module of Acrobat can be used to index documents. Typically, the size of the index to a PDF document is between 10 percent and 30 percent of the document itself.

A PDF document contains all of the information needed to render it exactly as it appeared to the creator. All Adobe Acrobat or PDF readers, which Adobe distributes as free software, will render a PDF encoded document in exactly the same way even though the software application used to create the document is not available. It is this universality of rendering without regard for hardware or software application domains that supports the portability of PDF documents.

Acrobat 3.01 supports several powerful navigation features. Hypertext links that have been added to other parts of a document or to external sources can be activated by clicking an icon. Acrobat Catalog can be used to generate indexes to one or more PDF documents and these indexes can be queried through Adobe Exchange to retrieve documents containing a specific word or phrase by using word stemming, "sounds like," Boolean expressions (and, or, and not), and proximity searching. In addition there are standard navigational features such as "browse" and "find. The browse feature includes several icons, such as "Display First Page," Display Next Page," "Display Previous Page, and the like that can be used to move sequentially through a document.

In Acrobat a PDF document can be revised either through deleting and adding pages or editing a single line of text. A PDF document can be made non-revisable by invoking a security option that prohibits any changes. This security option can be password protected against unauthorized removal of the restriction against changes so that PDF documents can not be edited or revised. This non-revisable feature of Adobe Acrobat would be particularly in protecting the integrity of published patents by ensuring that no changes or alterations could be made to them.

Acrobat does not support an export function so PDF documents can only be rendered within Acrobat. Nonetheless, a PDF file is in essence a Postscript file and with the appropriate software it can be converted to ASCII or RTF documents. One software package called Ghostscript contains a Postscript editor and interpreter that can be used to open and edit Postscript or PDF files. Revisions can be made and saved as ASCII, PDF or RTF documents. The other software package is called PDF2RFT and is available as a "plug and play" installation from BVAMYFRA, a software company in France. Consequently, even though Acrobat 3.01 does not support an export capability for PDF, records can still be converted with Ghostscript or PDF2RFT. Nevertheless, this is another instance of software dependence of electronic sources that militates against long-term access to them. Adobe currently supports backward compatibility from Acrobat 3.01 to 2.0. The company is likely to continue support in the future so in the short-run Acrobat viewers should be available to view electronic records embedded in PDF. However, in a ten to twenty year time frame and beyond, Acrobat support and readability become much more problematic.

Adobe PDF is a proprietary file format. Acrobat software, which includes the Distiller, Exchange, and Catalog discussed above, must be used to create, index, and search for PDF documents or records that can only be rendered with the Acrobat Reader. However, Adobe Systems makes the reader available free to anyone so that some people consider PDF a de facto industry standard. One implication of the wide spread availability of Acrobat readers at no cost is that anyone with a PC and access to an Internet service provider could easily retrieve and

render approved patent grants that the PTO published in PDF on its home page. This would require the PTO to convert approved patent grants from the file format(s) used for internal processing of electronic patent applications to PDF. However, the integrity of approved electronic patents could be protected against alteration by invoking the security option that blocks any change.

A survey of recent issues of the *Federal Register* and selected federal agencies that support a home page revealed that without exception any document that was available for viewing or downloading could be retrieved as a PDF document. There appear to be two primary reasons for this. First, the creator of a PDF document -- in this instance the federal agency -- controls how a PDF document looks when it is rendered on a monitor or printer. Second, PDF readers are widely available at no cost to users, which gives PDF a de facto universality.

It has been suggested that several federal agencies that require submission of information in electronic form are considering a requirement that the material be in PDF. The only documented instance of this requirement is in a 1997 draft proposed technical standards and guidelines for electronic filing for the United States Courts and it is unclear how binding this is. Several years ago there was a federal agency working group led by the Department of Defense to establish PDF as the standard file format for the removal of electronic records to the National Archives and Records Administration (NARA). However, NARA's response was "lukewarm" and eventually the working group disbanded. One very knowledgeable DOD staff member knows of no regulation mandating the use of PDF. It has been suggested that the Food and Drug Administration is considering the use of PDF in the submission of drug applications but this has not been confirmed.

STRENGTHS

- PDF documents are easily created
- PDF documents can include embedded spreadsheets, graphics, and text
- PDF documents can contain internal and external hypertext links
- PDF documents can be protected against revisions
- PDF Reader software is free
- PDF Acrobat software cost is modest
- PDF is already used by many Federal agencies in electronic publication/dissemination program
- Adobe Acrobat (PDF Writer, PDF Distiller, Catalog, and Exchange) and Reader run on multiple platforms
- Acrobat Exchange supports very powerful query and retrieval navigational tools
- Adobe intends to maintain backward compatibility across two generations of software
- The development of SVG as a W3C standard is compatible with PDF

WEAKNESSES

- PDF is a proprietary file format maintained by a single vendor
- PDF documents are not processible as such
- PDF documents can only be converted or migrated to another file format through the use of third part software (i.e., PDF2RTF)

6.1.4.1 Adobe Circulate

Adobe recently introduced a software tool called Adobe Circulate that incorporates three core Adobe technologies: Adobe Acrobat Exchange for search and retrieval, Adobe Capture (OCR Engine), and the Adobe PDF format. Adobe Circulate supports a capability that can combine or stack multiple files or individual pages into a single electronic file folder (e.g., a patent case file) in any specified order or sequence. Circulate also supports the export of PDF files into Microsoft Word or WordPerfect for editing, which eliminates the non-export drawback of Acrobat 3.01. It is likely that the storage requirements for PDF documents generated within Adobe Circulate will parallel those for PDF documents generated within Adobe Acrobat.

While Adobe Circulate includes several very powerful enhancements to Acrobat, it is nonetheless a proprietary product that Adobe is promoting in the high end of the client/server market. At this juncture it is unlikely that the proprietary character of Adobe Circulate will change in the foreseeable future. The proprietary aspect of Adobe Circulate along with its relatively high cost is a major drawback relative to its adoption as a file format in the submission of electronic patent applications and related electronic records.

STRENGTHS

- Adobe Circulate offers several very powerful tools for use with PDF encoded documents
- Adobe Circulate supports strong navigational tools, such as exact string matches, Boolean queries, and proximity queries
- Adobe Circulate supports the exporting of PDF encoded documents into conventional (e.g., Microsoft Word or WordPerfect) word processing applications for revisions, which helps to ensure long-term processibility
- The combination and stacking feature of Adobe Circulate would help support control of the internal content of electronic patent applications
- Acrobat Reader can read all PDF documents encoded or captured in Circulate
- Adobe Circulate runs on multiple platforms

WEAKNESSES

- PDF documents in Circulate are revisable
- Adobe Circulate is a proprietary product maintained by a single vendor
- The cost of the software is substantial

6.1.5 Rich Text Format

Rich Text Format (RTF) was developed by Microsoft as an "open format" for the interchange of Microsoft Word and other word processing software that retains all of the formatting instructions (how a document looks) such as bold, italics, type font size, color, and the like from the original document. RTF consists of a set of "common formatting commands" that the RTF writer substitutes for the formatting or markup instructions that word processing software generate. A RTF reader reads a RTF document and replaces the "common formatting commands" with their equivalents in a specific word processing package. For example, a document created in Microsoft Word 97 could be converted to RTF and then read by Word Perfect 6.0 running on Windows 3.1. In both Microsoft Word 97 and WordPerfect 8.0 this conversion is accomplished

by saving a document as RTF. Later, the RTF document can be converted to any word processing software package that supports RTF.
RTF is a proprietary product developed and maintained by Microsoft. Nonetheless, its wide spread availability and use means that it is a "de facto standard. As long as there is a market place requirement for the interchange of documents created by different word processing software running on different operating systems, it is likely that Microsoft will continue to support RTF. In the future, a new product that is better than RTF may become available and challenge or displace RTF. Market place considerations are likely to require that the new product have some form of backward compatibility with RTF. There is some risk, of course, that vendors may decide that here is insufficient pressure to provide backward compatibility. This risk appears to be minimal. Therefore, RTF may be a good option for the long-term storage of processible textual material.

STRENGTHS

- Encodes formatted text and graphics for interchange across multiple word processing applications
- Supports hyperlinks to external material such as spreadsheets and images
- PDF documents can be converted to RTF and imported into Microsoft Word or Corel WordPerfect 8
- RTF software is a standard utility in Microsoft Word and Corel Word Perfect 8
- Converting a word processing document to RTF is simple and involves nothing more than clicking on a "save as RTF" command

WEAKNESSES

- Revisable text with no internal evidence of change that can raise questions about the integrity of electronic patent applications
- RTF is a proprietary file format with a single vendor source
- RTF does not lend itself to multimedia

6.1.6 ISO/IEC 8211

In 1985 ISO 8211 was approved as an international standard for data interchange and "archiving" that was content-independent and medium-independent. In 1994 ISO 8211 was updated and re-issued as ISO/IEC 8211, Information Technology - Specification for a Data Descriptive File for Information Interchange.
ISO/IEC 8211 enables the development of general purpose software that can create a data stream consisting of a standardized data description and data that can then be moved into a target system where ISO 8211 conforming software desegregates the bit stream and interprets it according to its data description. This standard is intended to support the direct interchange of any kind of digital material - ASCII character code, vector graphics, bit map images, spreadsheets, and databases, among others - to a known target system without any loss in content. In addition, this standard supports the creation of ISO 8211 conforming bit streams for which the target system is unknown. This is particularly appealing to institutions with the mandate to preserve long-term access to electronic material but do not know what the target

system will be in the future. For ISO/I/EC 8211 to work both the creating and target systems must have ISO/IEC 8211 conforming software. A crucial aspect of ISO/IEC 8211 is that it is not necessary for implementing software to understand the internal operations of the originating system or the target system.
Practically speaking, ISO/IEC 8211 conforming software comprises a "black box" that appears to make it ideal for the PTO to support long-term access to electronic patent applications. Unfortunately, ISO 8211 was perceived as meeting a very narrow need and very little software was ever developed. Alfred A. Brooks, the chief architect of ISO/IEC 8211, wrote several software implementations in Fortran and C in the late 1980s and early 1990s and there were several other efforts that never fully materialized, but there never was a commercial software implementation by a major vendor. Today, ISO/IEC 8211 serves a niche market through incorporation into the American National Standards Institute (ANSI) draft Spatial Data Transfer Standard (SDTS). There is a software implementation of ISO/IEC 8211 under development by the United States Geological Survey.
Despite its intuitive appeal and attraction in terms of supporting long-term access to electronic patent applications, ISO/IEC 8211 is not a mainstream technology application and the PTO should not consider its use.

STRENGTHS

- ISO/IEC is a non-proprietary national and international standard
- ISO/IEC 8211 can provide maximum portability of electronic material, particularly when the target system is not know

WEAKNESSES

- There is virtually no market penetration of ISO/IEC 8211 software implementations

6.1.7 ISO 8613 (ODA/ODIF)

In the mid 1980s there was a concerted effort within the international information technology community to develop standards and protocols that would support the concept of "open systems," which consist of components with published interfaces that an be used to connect them with other components without obstacles from operating systems or application constraints. A major step in supporting open systems was acceptance of the Open Systems Interconnection (OSI) seven-layer architectural model as an international standard in 1986. One of the OSI follow-on activities was the development of the Office Document Architecture/Office Document Interchange Format. Known officially as ISO 8613, ODS/ODIF was approved in 1988.

The rationale for ODA/ODIF was to facilitate the exchange of office documents such as reports, letters, memoranda, and the like between dissimilar computer platforms and applications without any loss of content. ODA is a robust standard that defines a business document architecture in terms of content and two hierarchical structures: a logical structure and a layout structure. It is noteworthy that the separation of the logical structure of a document from its layout structure parallels that of SGML. In addition, ODA supports character data and graphic data. A document encoded as character data is intended to be revisable while graphic data such as a bit map image of the same document is intended to be non-revisable. ODIF defines the encoding of a bit stream that is to be transferred from one system to another.

At the time ODA/ODIF was approved many knowledgeable observers thought that it was the standard with the greatest potential for electronic office records. In order to "jump start" the development of software implementations of ODA/ODIF the National Archives of Canada sponsored a pilot ODA/ODIF demonstration project. Similar software implementations were begun in Europe. Although these demonstration projects were successful, there was very little follow-on interest in ODA/ODIF. The primary reason for this was the market place acceptance of Standard Generalized Markup Language (SGML), a competing ISO generalized markup language that had been approved in 1986. Consequently, no software implementation of ODA/ODIF has penetrated the market place. ODA/ODIF has no potential use in the PTO life cycle management of electronic patent applications.

STRENGTHS

- Approved international standard
- Combines revisable and non-revisable text in the same document package
- Potential for providing an envelope or wrapper for multimedia material

WEAKNESSES

- Virtually no market place presence in terms of software implementations

6.2 VECTOR GRAPHICS

Vector graphics consists of mathematical descriptions of one or more image elements, which are used by the rendering application to construct a final image. At its simplest level each of these image elements may consist of line segments or drawings that typically are defined as a shape consisting of a starting point, a direction, and a length. Straight and curved lines can be combined to form other, more complex geometrical objects. Vector graphics are used largely in digital cartography, Geographic Information Systems (GIS), and Computer-Aided Design/Computer-Aided Manufacturing (CAD/CAM) applications. In vector graphics applications, lines and shapes are associated with information that specifies size, shape, position relative to the overall image, color, and other attributes. Unlike bit map images (discussed later),

vector graphics are processible as ASCII encoded data, which, of course, means that alterations and modifications are easily made.

6.2.1 Initial Graphics Exchange Specification (IGES)

In the late 1970s users of vector graphics data files became interested in having a common format for the "platform-independent" interchange of vector data between CAD/CAM systems and other vector oriented applications. Version 1 of a standard called the Initial Graphics Exchange Specification (IGES) was adopted in 1981as an American National Standard. The most recent version of IGES is 5.2. Although IGES is a non-proprietary standard, its complexity is such that it tends to be implemented only in fairly high-cost CAD/CAM applications such as those used in the automobile and aerospace industries. Typically, IGES export and import functions are standard utilities in CAD systems.

STRENGTHS

- IGES is a non-proprietary standard
- IGES supports the exchange of digital data between CAD applications
- IGES can represent many different types of vector representations of geometric entities
- There are a substantial number of IGES software implementations

WEAKNESSES

- IGES is extremely complex
- IGES is used by a relatively small number of typically large entities

6.2.2 Computer Graphics Metafile (CGM)

Computer Graphics Metafile (CGM), which is also known as ISO 8632, is an international standard for the exchange of vector data between multiple computer platforms and software applications. Developed in 1986 under the auspices of the International Standards Organization (ISO) and the American Standards National Institute (ANSI), CGM was revised in 1992 and included a minimal level of implementation. This minimal level of implementation was intended to address an incompatibility between filters caused by some implementations of the 1986 version that ignored some features of CGM. In order to further the full exchange of CGM vector data, the National Institute of Standards and Technology implemented a testing service that verifies conformance of CGM export and import filters to the 1992 minimal level of implementation.
CGM filters are widely used in business graphics and word processing applications, such as Microsoft Office 97. Many of these filters (e.g., Microsoft Office 97) are certified by NIST as being conformance with ISO 8632: 1992.

STRENGTHS

- CGM is a non-proprietary exchange standard
- CGM filters run on multiple computer technology platforms

- CGM is widely implemented through standard utility filters for importing and exporting vector material
- NIST conformance testing of CGM filters
- When installed as a standard utility, a CGM filter is easy to use and requires no special skill or knowledge

WEAKNESSES

- CGM filters developed before 1992 may not support universal interchangeability of vector files
- CGM is not a mainstream W3C supported standard

6.3 BIT MAP IMAGES

Another form of digital representation called the graphical image or bit map image, is a numerical representation of the variation in monochromatic or chromatic reflectance of a target area such as a photographic image or a page of paper. This variation in reflectance is captured at what is called the picture element level (abbreviated as pixel), which typically is defined as a specified number of dots per inch (dpi). The number of dots per inch measured horizontally and vertically determines pixel size. For example, a resolution of 200 dpi means that each inch of vertical space is divided into 200 horizontal lines with each line divided into 200 dots per inch. For an 8.5 by 11 inch page, this would mean there are 2200 lines (11 inches X 200 dpi) and each line would consist of 1700 pixels (8.5 X 200 dpi) or dots per line. The resolution of detail for a given image is measured by the number of pixels or dots per inch (dpi), which generally ranges from 100 to 600. The higher the number of dots per inch the higher the detail or resolution of an image. Of course, the higher the number of dots per inch, the greater the size of the scanned file. The storage requirements for bit map images are substantial so techniques for reducing them, called compression, have been developed.

6.3.1 Compression Techniques

The function of compression techniques is to reduce the volume of data being stored or transferred and to reconstruct the full image for display purposes. Compression techniques are characterized as lossy or lossless. Typically, lossy compression techniques discard data based upon the limitations of human vision. For example, human vision can only process about 10,000 different colors simultaneously, and differences in color are most easily distinguished when the contrast is significant. Thus, lossy compression techniques discard data that the brain does not use. Once bits or data are discarded it is impossible to reconstruct fully the original image. In contrast, lossless compression techniques do not discard any data; the decompression stage produces exactly the same data read at the prior to compression. As a general rule, lossless compression techniques can help support the integrity of reconstructed bit map images.

6.3.1.1 Group 4 CCITT Compression

The International Telecommunications Union or ITU (previously known as the International Telegraph and Telephone Consultative Committee (CCITT) is a United Nations standards organization that has developed protocols for the transmission of bi-tonal (black/white) images over telephone lines and data communication links. One of these protocols or standards for encoding bi-tonal or one-bit images is called Group 4, which was developed specifically for bit map images stored on disks and transmitted across networks. It is a Run Length Encoding (RLE) lossless compression technique that can achieve compression ratios on the order of 15:1 for printed, typed, and handwritten material. Group 4 compression is implemented in most digital image software applications currently available.

STRENGTHS

- Lossless compression
- International standard
- Widely implemented

WEAKNESSES

- Somewhat slow execution rate
- Limited to bi-tonal images

6.3.1.2 Joint Bi-level Image Group (JBIG) Compression

Joint Bi-level Image Group Compression is an international standard for the lossless compression of bi-tonal images developed by the Joint Bi-Level Image Group, a standards committee of the International Standards Organization. The rationale for its development was to develop an "adaptive compression technique" that would be more efficient than Group 4. Indeed, JBIG achieves compression ratios that are between 1.1 and 1.5 greater than Group 4. This gain in efficiency must be balanced against the fact that JBIG contains a number of patented processes, the most prominent of which is called the "arithmetic Q-code," owned by IBM. Although JBIG was developed for bi-tonal images, it can be used to compress color and gray-scale images with a "bit depth" of 8 bits (256 pixels). This is particularly relevant where lossless compression is required. Despite JBIG status as an international standard, there have been very few software implementations and, as such, market penetration is minimal.

STRENGTHS

- International Standard
- Highly efficient
- Lossless compression
- Can be used (with limitations) with gray scale and color images

WEAKNESSES

- Part of the compression technique utilizes a proprietary arithmetic coder
- Very little market penetration

6.3.1.3 Portable Network Graphics (PNG).

PNG uses a variation of the LZ77 lossless compression algorithm developed by Phil Katz that is formally called “Deflate,” but is known more generally as "pkzip" (pronounced P-K-zip). The development of PNG was promoted by the World Wide Web (W3) Consortium for use on the Internet as a replacement for the Graphics Interchange Format (GIF) developed by CompuServe. PNG offers several features not available in GIF, which include true color images up to forty-eight bits per pixel, gray scale images of up to 16 bits per pixel, detection of file corruption, image gamma information (color), 100 percent lossless compression, hardware and platform independence, and ease of future extensibility. And, a version of LZ77 is in the public domain.

STRENGTHS

- Lossless compression
- Non-proprietary compression algorithm
- True color
- Highly portable

WEAKNESSES

- Very few software implementations so market penetration is limited
- Persistence over time is undetermined

6.3.1.4 Graphics Interchange Format (GIF)

The Graphics Interchange Format, which was developed by CompuServe for on-line transmission of lossless bit map images (bi-tonal, gray scale, or color), employs the LZW compression technique, which, as noted below, is a proprietary product of the Unisys Corporation. When GIF was developed it was widely believed that LZW was in the public domain. The fact that GIF is not in the public domain is only one of several problems. GIF employs no error correction capability. GIF is limited to 256 values for a single pixel, which works fine with bi-tonal or gray scale tone images but not with full color.

STRENGTHS

- Lossless compression
- Accommodates bi-tonal, gray scale, and color images (with limitations)
- Supported by CompuServe

WEAKNESSES

- Not in the public domain
- Lacks error correction capability
- Limited color values

6.3.1.5 Lempel-Ziv-Welch

There are several variations of LZW, which is a lossless compression technique. The first is called LZ77, and was created in 1977 by Abraham Lempel and Jacob Ziv (hence LZ). One year later they developed an enhanced compression scheme called LZ78, which the Sperry Corporation (now part of the Unisys Corporation) subsequently patented. In 1985 Terry Welch modified the LZ78 compression scheme which the Sperry Corporation patented it as LZW. Unlike Run Length Encoding (RLE) compression, the compression technique used in LZ78 or LZW is adaptive because it adapts or adjusts to the specific characteristics of any bit map image and dynamically encodes the bits in order to maximize efficiency.

STRENGTHS

- Lossless compression
- Widely implemented
- Highly efficient algorithm

WEAKNESSES

- LZ78 and LZW are proprietary compression algorithms

6.3.1.6 Joint Photographic Experts Group (JPEG)

JPEG is the product of an expert group organized by the International Standards Organization to develop a standard for compressing either full-color or gray-scale digital images and rendering these images for human viewing rather than by computer viewing. Consequently, JPEG is inherently lossy, though the amount of information lost from any image can usually be set by the user and, if done properly, may not be easily detectable by the human eye. JPEG divides an image into units of 8 by 8 pixels and mathematically transforms the value of each unit pixel into a cosine function. On average, less than 10 percent of the cosine functions in any 8 X 8 pixel unit are required in rendering an accurate representation for the human eye. If further degradation of the image is permitted (i.e., thumbnail sketches), then even fewer cosine functions are required, and this results in even greater compression. There are many software implementations of JPEG, including public-domain software. Because JPEG is a lossy compression algorithm, which potentially can undermine the integrity of electronic patent records that contain bit map images, the PTO should not permit its use.

STRENGTHS

- International standard
- Very efficient compression
- Widely implemented

WEAKNESSES

- Lossy compression
- Contains proprietary components

6.3.1.7 Motion Picture Experts Group (MPEG)

MPEG is an international standard (like JBIG and JPEG) that was developed for the compression of bit streams that contain both moving images and audio or sound signals. It is used widely for the storage of audio and visual data on CD-ROMS and a variety of multimedia environments. In MPEG implementations of moving images, JPEG compression techniques are used for individual frames and other lossy techniques are used to compress data between frames. Currently, MPEG is the only non-proprietary compression algorithm available for moving images and audio so the PTO has no choice about a lossy versus lossless compression algorithm, short of storing uncompressed moving images and audio. This is not a viable alternative in most instances because of the enormous storage requirements it would entail.

STRENGTHS

- International standard
- Widely implemented
- Substantial market penetration so persistence over time is likely

WEAKNESSES

- Lossy compression algorithm

6.4 IMAGE FILE HEADER

Bit map images carry no inherent intelligence. Hence, for a computer to process bit map images certain information must be provided in what is called an image file format that typically consists of an file header and one or bit map images. An Image File Header (IFH) may store selected attributes of a graphics image such as its height and width, the bit depth (bi-tonal, gray scale, and color), the byte order (whether byte order is to be read from left to right or right to left), the compression technique used, the scanning resolution used, a pointer to where the bit map image data begins, and in some instances a pointer to an Image File Directory (IFD) that contains more detailed information.

Image file formats often are proprietary products supplied as part of an integrated digital imaging system. Proprietary image file formats should be avoided because they require access to

proprietary software, which militates against long-term access. Currently, no non-proprietary or international standard exists for image file formats. Nonetheless, the Tagged Image File Format (TIFF) is widely implemented and consequently can be considered a de facto standard.

6.4.1 Tag Image File Format (TIFF)

TIFF was developed by the Aldus Corporation for storing black and white images created by scanners and desktop publishing applications. It has gone through several revisions and extensions since its release in 1986, the most recent being TIFF 6.0, which was released in 1992. Today, TIFF can be found in many digital imaging and desktop publishing software packages. TIFF files are organized into three sections: (1) the Image File Header (IFH), (2) the Image File Directory, and (3) the bit map data. The Image File Header is always the first eight bytes of a TIFF file and contains three items of information: (1) the byte order, (2) the TIFF version, and (3) a pointer to "offset" where the first Image File Directory (IFD) begins. The Image File Directory in TIFF contains detailed information about each bit map image organized by tags into fields. TIFF 6.0 identifies four "baseline" images: (1) bi-tonal (black and white), (2) gray scale, (3) palette color, and (4) full color. Each baseline has a minimum set of tags or fields that include Image Width and Length, BitsPerSample, Compression, Resolution, StripOffSets, RowsPerStrip, among others. As noted earlier, the third section of a TIFF file is the actual bit map data.

TIFF 6.0 supports Run Length Encoding, Group 4, LZW, and JPEG compression techniques, although there is some question about the effectiveness of the JPEG implementation. Version 6.0 of TIFF does not support the use of JBIG or PNG compression algorithms that are more appropriate for PTO storage and transfer of electronic patent applications containing bit-map images.

Part of the rationale for the creation of TIFF was to ensure the interchange of bit map images. However, over time so many versions and "flavors" of TIFF have been implemented that it is not always possible for one TIFF software implementation to read correctly a TIFF file created by another TIFF software implementation. This is a serious problem with regard to long-term access to bit map images. However, this problem could be mitigated somewhat if the PTO were to issue its own TIFF specifications for electronic patent applications that producers of patent applications would be required to use. This would ensure uniformity in image file headers used in electronic patent applications and facilitate long-term access to the material.

STRENGTHS

- A de facto industry standard
- Widely implemented with major market penetration
- Can be customized to serve PTO needs
- Highly portable within conforming implementations
- Persistence over time is likely

WEAKNESSES

- There are various "flavors" to TIFF in use that can create readability problems

6.5 SUMMARY COMPARISON

This assessment of these seventeen (17) file formats is summarized in two tables. Table 1. deals with the file formats based on the criteria of applicant process, while Table 2. addresses the file formats based on the criteria developed for PTO processes.

Table 1. APPLICANT PROCESS CRITERIA							
File Format	**Multi-Media**	**Navigability**	**Portability**	**Integrity**	**Confidentiality**	**Rendering**	**Storage**
SGML	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Nominal**
XML	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Nominal**
PDF	**Yes**	**Yes**	**Yes**	**Partial**	**No**	**Yes**	**Nominal**
RTF	**No**	**No**	**Yes**	**No**	**No**	**Partial**	**Increase**
ISO 8211	**Yes**	**No**	**Yes**	**No**	**No**	**No**	**Nominal**
ISO 8613	**Yes**	**No**	**Yes**	**No**	**No**	**No**	**Nominal**
IGES	**NA***	**No**	**Yes**	**No**	**No**	**Yes**	**Nominal**
CGM	**NA**	**No**	**Yes**	**No**	**No**	**Yes**	**Nominal**
SVG	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Nominal**
LZW	**NA**	**No**	**NA**	**Yes**	**No**	**Yes**	**No**
Group 4	**NA**	**No**	**NA**	**Yes**	**No**	**Yes**	**No**
JBIG	**NA**	**No**	**NA**	**Yes**	**No**	**Yes**	**No**
JPEG	**NA**	**No**	**NA**	**No**	**No**	**Yes**	**No**
MPEG	**NA**	**No**	**NA**	**No**	**No**	**Yes**	**No**
GIF	**NA**	**No**	**Yes**	**Yes**	**No**	**No**	**No**
PNG	**NA**	**No**	**Yes**	**Yes**	**No**	**No**	**No**
TIFF	**NA**	**No**	**Yes**	**NA**	**No**	**No**	

*Not Applicable

Table 2. PTO PROCESS EVALUATION CRITERIA								
File Format	**Multi-Media**	**Navigability**	**Portability**	**Integrity**	**Confidentiality**	**Issuance**	**Persistence**	**Storage**

SGML	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Yes**	**Nominal**
XML	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**Yes**	**Nominal**
PDF	**Yes**	**Yes**	**Yes**	**Partial**	**No**	**Yes**	**Yes**	**Nominal**
RTF	**No**	**No**	**Yes**	**No**	**No**	**No**	**Yes**	**Increase**
ISO 8211	**Yes**	**No**	**Yes**	**No**	**No**	**No**	**No**	**Nominal**
ISO 8613	**Yes**	**No**	**No**	**No**	**No**	**No**	**No**	**Nominal**
IGES	**NA***	**No**	**Yes**	**No**	**No**	**Yes**	**Yes**	**Nominal**
CGM	**NA**	**No**	**Yes**	**No**	**No**	**Yes**	**Yes**	**Nominal**
SVG	**Yes**	**Yes**	**Yes**	**No**	**No**	**Yes**	**?**	**Nominal**
LZW	**NA**	**NA**	**No**	**Yes**	**No**	**Yes**	**Partial**	**No**
Group 4	**NA**	**NA**	**No**	**Yes**	**No**	**Yes**	**Partial**	**No**
JBIG	**NA**	**NA**	**No**	**Yes**	**No**	**Yes**	**Partial**	**No**
JPEG	**NA**	**NA**	**No**	**No**	**No**	**No**	**Yes**	**No**
MPEG	**NA**	**NA**	**No**	**No**	**No**	**No**	**Yes**	**No**
GIF	**NA**	**NA**	**No**	**Yes**	**No**	**No**	**No**	**No**
PNG	**NA**	**NA**	**No**	**Yes**	**No**	**No**	**Partial**	**No**
TIFF	**NA**	**NA**	**Yes**	**Yes**	**NA**	**NA**	**Yes**	**No**

*Not Applicable

7 ANALYSIS AND RECOMMENDATIONS

The anticipated economies and efficiencies to be derived from the implementation of the electronic patent application processing can be achieved only if the process is standardized so that whatever file formats applicants submit can be readily received, processed and maintained for the full retention life by the PTO. Achieving this level of standardization is likely to involve a substantial investment in technology and the acquisition of the skill and knowledge to utilize the technology. A fundamental question for the PTO is how this cost should be allocated? Should the PTO issue a set of requirements defining a template that uses a PTO defined "standard" file format for text, bit map images, and vector graphics that third party software vendors would then write programs to support? This has the net effect of shifting a considerable part of the cost of implementing the electronic patent application program to the patent applicants. Or should the PTO issue very broad guidelines that give patent applicants considerable flexibility in the use of software and file formats with the expectation that the PTO would bear all of the cost of moving electronic patent case files into a standard file format? Another alternative would be for the PTO to issue requirements and specifications that standardize the file formats. Then the PTO and independent software vendors could develop electronic records creation and submission applications that implements the file formats and make the software available to anyone submitting an electronic patent application.

The PTO's response to this question of cost allocation depends in part upon which file format is adopted as the "standard." If PDF were to be adopted then the cost of purchasing Acrobat is modest and it would not be unreasonable for the PTO to require patent applicants to bear this cost. Using Adobe Acrobat for accessing PDF files would also help ensure the integrity of applications during the transmission and receipt stages as well as during the publication and maintenance phases. Acrobat Catalog and Exchange support very powerful navigation tools that might be of some interest to patent applicants but would be of considerable interest to patent examiners. However, PDF also is a proprietary product that could makes its use over an extended period of time problematic. Furthermore, the ability to transfer PDF electronic patent applications to a new technology environment is constrained. Of course, if Acrobat Capture and Circulate were used this would be less of a problem, but would require additional software outlays by the applicant and the PTO. This benefit has to be weighed against the projected costs of acquiring this software for PTO examiners.

If SGML were adopted as the standard, then the software, particularly the patent preparation software, is considerably more expensive and difficult to use. SGML software is more expensive than Acrobat and more difficult for users to master. However, SGML-defined electronic patent applications could be navigated easily and annotated for purposes of review, examination and publication. SGML could also be easily migrated to a new technology. However, there is no inherent functionality in SGML that supports document integrity.

Use of XML and the Scaleable Vector Graphics (SVG) extension to prepare and process electronic patent applications can have considerable potential for the PTO. According to a senior Adobe official, PDF should be compatible with XML and the SVG extension. If in fact

this turns out to be the case, then it should be relatively easily to move PDF electronic patent applications into XML.SVG or vice versa. This would give the PTO the best of the PDF and the XML worlds. There is very little XML software in place now but it appears that JAVA can be used to develop relatively inexpensive XML software.

The U.S. Federal Courts of New Mexico adopted the position that the best way to standardize the submission of electronic pleadings would be to develop prototype XML software that accepted PDF documents. The introduction of Precision Graphics Meta Language (PGML) by Adobe and its W3C successor, SVG, is seen as a significant development that could achieve the desired standardization at a minimum cost to the PTO and applicants.

There are tradeoffs between the file formats reviewed in this study and, as such, there is no single file format that satisfies fully all of the criteria. On balance, XML and its extensions, including the newly proposed SVG extension for graphic representations, appears to best satisfy the criteria except for one major problem, namely, the SVG extension is not likely to be adopted for over a year and currently there is very little XML software available. All of the evidence suggests that, as predicted by its proponents, XML and the SVG extension (as well as other XML extensions) will rapidly evolve to become the preferred file format for conducting electronic commerce. It is also assumed that XML will become as ubiquitous as HTML in terms of creating XML output from desktop office productivity applications, send/receive e-mail in XML format, and be integrated (at no or low cost) as a viewing/navigation/reproduction capability into Internet browsers and desktop productivity applications. Based on the these overall considerations and the detailed assessment, the following recommendations are made:

- **Provisionally adopt Extensible Markup Language (XML) and its extensions, including Scalable Vector Graphics (SVG), for supporting electronic patent application preparation, review and examination.**

  The provisional nature of this recommendation is due to the current status of XML and its extensions, in that full-scale implementation support has not been delivered and, as such, widespread user acceptance has not yet been established. It is expected that full vendor and user support will materialize over the next one to two years whereupon the provisional nature of this recommendation would be lifted.

  This recommendation is based on the ability of XML to support structural tagging and navigation and the ability of SVG, along with other XML extensions to provide support of graphics and more complex work units. Since XML can also recognize “well-formed documents”, non-tagged electronic records such as e-mail (that is not transmitted in XML) should also be able to be cognized and processed. XML and its associated extensions and capabilities (as currently defined or under definition) meets the majority of the assessment criteria, including those related to preparation, rendering, portability, multi-media, navigability and, presumable, format persistence. As such, XML should provide a single file format that can create, store and accurately render the content and structure of all electronic records contained in a patent case file. Potential drawbacks of this format may be the cost and ease of preparation for the applicant and the lack of any inherent preservation of integrity. In support of this recommendation, it is suggested that the PTO:

- Develop an XML Document Type Definition that could be implemented and include the SVG extension so as to standardize the submission of electronic patents.
- Review the XCI project of the New Mexico Federal Courts and learn how XML with the SVG extension is being implemented.
- Develop and prototype software specifications to support the creation of XML with the SVG extension for electronic patent applications and related documents.

- **Provisionally adopt Extensible Markup Language (XML) and its extensions, including Scalable Vector Graphics (SVG), for electronic publication and dissemination of granted patents.**

  Since the as-filed and amended, if required, patent application and associated electronic records are recommended to be produced and stored in XML (with extensions) format, publishing of the patent in XML is the logical and cost-effective choice.

  One drawback with XML, as with SGML, is that the file format does not provide an inherent ability for ensuring the integrity of content and structure. If the requirement for providing inherent file integrity (non-revisable) of the published patent is deemed essential, and if a non-revisable method for rendering the published patent is not possible with XML (and the SVG extension), then PDF would be a better choice. PDF has the obvious disadvantage of being a proprietary standard and may be superceded over time by the .SVG extension of XML, particularly in Internet-based electronic commerce.

- **Provisionally adopt Extensible Markup Language (XML) and its extensions, including Scalable Vector Graphics (SVG), for long term maintenance and preservation of the electronic records in the patent case file for the full retention life.**

  Assuming that the XML and its extensions become the industry standard for electronic records produced in electronic commerce, the retaining the electronic patent case file records XML should meet the long term maintenance, use and preservation requirements related to format persistence, rendering and portability.

  Since XML does not currently provide an inherent ability for ensuring the integrity of content and structure, other measures must be taken by the PTO to meet this requirement.

- **Adopt Scalable Vector Graphics (SVG) as the standard for vector graphics.**
- **Employ only lossless compression methods for bit map images, including Group 4 ITU and JBIG for bi-tonal images, and PNG for grayscale and color images**
- **Define a "standard" TIFF image header (including any PTO-specific tags) and require the producers of electronic patent applications to use it when the submission includes bit map images**