Scoping Document on PDF Optimisation

This document explores potential issues that users in the developing world might face in using websites that host PDF documents, particularly regarding the downloading and usage of research papers as PDF documents.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Table of Contents

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Glossary

Term	*Definition*
Cache	A collection of data previously downloaded from elsewhere that is stored locally for some amount of time, allowing fast access for duplicate requests.
CSS	Cascading Style Sheets
Firewall	A piece of hardware or software that prevents unauthorised connections being made in a network environment.
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
HTTP Resuming	Functionality on a server allowing a download over HTTP to restart from the last point reached if the connection is broken mid-download.
ISP	Internet Service Provider
kB	Kilo Bytes
kbps	Kilo bits per second
LAN	Local Area Network
Malware	Malicious software: software designed to install itself onto and / or damage a user's computer without consent.
Metadata	Information that describes another set of data
OCR	Optical Character Recognition
PDF	Portable Document Format
Proxy	A network service allowing computers to connect to another service indirectly.
XML	eXtensible Markup Language

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Introduction

This document explores potential issues that users in the developing world might face in downloading and using PDF documents from the Internet, particularly in the context of academic research.

The Portable Document Format (PDF) can be an efficient way of publishing content for reading on screen or in print. However, some PDF documents are very large. These will be difficult for users in most developing countries to download.

We start by considering the format of information available and the limitations faced by potential users, such as the computing facilities and technical support available.

From this we go on to discuss in more technical depth the issues that users will face when accessing PDF documents. We consider existing solutions to these problems, propose possible new solutions and examine the effectiveness of both current and potential solutions.

Finally, we give conclusions and our recommendations for the future development of such websites. These include:

- Work closely with users to identify their expectations and intended use of the site, and any problems they experience;
- Ensure that development of the site is driven by user needs;
- Provide client-side tools to improve the efficiency and reliability of searching for and downloading files;
- Ensure that the site, and the documents offered for download, are optimised for the needs of low bandwidth users;
- Develop guidelines and tools to assist authors in creating efficient PDF documents.

Objective

The objective of this document is to recommend how to optimise access from the developing world to knowledge resources in the form of published documents.

Existing projects that have created websites aimed at users in the developing world have found that many of their users (and potential users) have been frustrated by the length of time taken to access information, which depends on:

- the size of documents to be downloaded;
- the speed and quality of Internet connections; and
- the interface to find and access documents.

Slow and intermittent access to information limits the number of users who find such systems practically usable, thereby reducing the effectiveness of these projects. In order to meet its objective, this document will identify methods of overcoming these factors, such as:

- methods of producing smaller PDF documents and
- download and bandwidth management.
- the website user interface;
- the search interface to material contained on the website;
- the requirements of users viewing and working with PDF documents after download.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Context

The PDF Standard

Portable Document Format (PDF) is a file format developed by Adobe Systems for representing documents in a device and resolution independent manner. Unlike other on-line formats, such as HTML, a PDF file can be reasonably expected to look exactly the same to every viewer, and to produce the same printed document.

These considerations have led many authors to publish on-line content as PDF, particularly for documents that are primarily formatted for printing. In fact, most documents stored on-line that are originally authored with a desktop publishing package or a word processor (such as Microsoft Word) are stored as PDF documents. The free availability of PDF viewing software makes it possible for anyone to view documents published as PDFs. According to the International Organisation for Standardization, an estimated 9.2% of the total material on the Internet is comprised of PDF documents.

The widespread use of PDF is often criticised on grounds of accessibility. Viewing HTML with a web browser allows the user to adjust font size, colour and other display properties to make the text easier to read. This is not possible with PDF. PDF documents are generally significantly larger than the same information stored as HTML and image files. Also, it is quite possible that a user will not have PDF viewing software installed, and would need to download this in order to view PDF documents. Users with slow or unreliable Internet connections may find that downloading viewing software is problematic.

Possible alternatives to the PDF standard are under development. These include the Metro format being developed by Microsoft and the OASIS Open Document Format for Office Applications. The PDF-Archive standard, currently in draft form, may also become commonly used for the long-term storage of PDF documents. It will contain a subset of the PDF 1.4 standard, selected for size optimisation. In the longer term, Digital Rights Management and Trusted Computing may also affect the accessibility of on-line documents.

Access to Knowledge Sources from Developing World Institutions

Many websites wish to provide information to users in the developing world, who often have limited bandwidth and intermittent connectivity to the Internet. It is therefore essential that the process of accessing the information contained in PDF documents through such websites is made as easy as possible, with the needs of developing world users in mind.

There are many potential usage scenarios for on-line PDF documentation. It is important to consider the different current and potential user and institution profiles when addressing access issues. These may include students, researchers, employees and volunteers at non-governmental organisations (NGOs), in-country staff, and information intermediaries such as librarians. While some users require the complete document for printing purposes, others will be satisfied with a compressed version that is faster to access.

Within an institution, the level of technical skills available; the network infrastructure; the level of access currently available and the number of users who will be downloading information all affect the appropriateness of different approaches to improving access.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Aidworld has carried out fieldwork in developing countries including Ethiopia, Kenya and Haiti. We have seen large variation between institutions visited in terms of:

- the bandwidth of connections;
- the network infrastructure available;
- the number of users sharing the connection and the infrastructure;
- the technical skills and support available;
- the level of network administration; and
- the ability to effect server or network-wide changes such as proxies or caches.

This variation means that it is essential to develop an understanding of user requirements, in terms of the most common and most severe impediments to access, in order to most effectively target appropriate solutions.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Issues

Users in the developing world often have to overcome a number of constraints to access information on the Internet. Reaching these users represents a major challenge.

To identify difficulties that will affect these users, we look at the tasks they may have to perform to get the information they seek from a document repository website.

Network Infrastructure

A major problem underpinning all tasks involving Internet access is the high cost of Internet connectivity in the developing world. Therefore, the average bandwidth available to users is very much lower than that in the developed world. For example, at the Malawi College of Medicine, each user has access to 0.5 kbps on average. This compares to 512 kbps or more for a broadband connection in the UK, and the user experience of university connections is regularly much faster than this. This throttling of connectivity affects everything the user does on the Internet: documents take a thousand times longer to download, even a web page may take up to ten minutes, or fail to load at all.

An additional problem is the reliability of the connection. Network connections in developing countries are frequently interrupted. It can often take multiple attempts to complete a download, especially for larger files. If breaks in connectivity tend to occur more frequently than the time required to download the entire PDF document, it is possible that a download will not be practically achievable.

The level of funding and skills available for network management means that the existing resources in many institutions are used very inefficiently. Currently, many networks with shared connections do not attempt to allocate bandwidth fairly between users, and do not attempt to block unauthorised use of the network or inappropriate content. Networks with shared connections are often flooded by the actions of computer viruses and other malware, and by unauthorised and inappropriate use of the network such as file sharing software.

It is not uncommon for web-based email services (web mail) to be used as an alternative to local email services due to either a lack of local email provision, a lack of awareness on the part of users, or a lack of trust in the reliability of local email services. This is a very inefficient use of bandwidth, particularly with popular bandwidth-heavy web mail sites such as Hotmail and Yahoo Mail. A combination of some or all of these factors can result in users experiencing unnecessarily slow or unusable connections.

Site Navigation

Users in developing countries may have lower levels of computer and Internet skills than those found in the developed world, and may not have English as their first language. If the site is inappropriately designed, or lacks language facilities, then users may be unable to find the information they require.

Some common website design issues are:

- unclear site structure and navigation;
- too many clicks to reach the desired information;
- use of pop up windows, making navigation difficult;
- large page sizes;
- poor search (see next section).

Users with older browsers or with browsers optimised for low bandwidths may not be able to view graphics or Javascript on websites.

Some technical issues that hinder access are:

- use of images without alternative text;
- use of image maps for navigation;
- Javascript required for navigation;
- use of Flash animations.

In addition to not being compatible with older browsers, these features increase the size of the page, further exacerbating the bandwidth issue.

Searching

The user may have problems with the search process that prevent them from finding relevant documents. These may include:

- search functionality is difficult to find;
- the search interface is difficult to use;
- the advanced search interface is not very powerful.

As a result, the search may fail to find the document they want, or produce too many documents. Other problems with the search results may include:

- slow search;
- large results page that takes a long time to download;
- search results contain too little information for the user to determine whether they are relevant;
- documents not being included in the search results because they are scanned images of the original document rather than plain text, and contain insufficient or no metadata;
- search results contain no indication of how long they will take to download.

Downloading

Downloading will always take some time on a low bandwidth connection, as PDF documents are relatively large. Even if some of the information in a PDF document is irrelevant to the user, they will still have to download the entire document.

Downloads may fail due to an unreliable connection, in which case they will need to be resumed or restarted, either by the user or by download management software. Larger files are more likely to fail to download due to the increased amount of time spent downloading.

Downloading during peak times is often slower and less reliable, and may cost more.

The following table shows minimum download time against bandwidth for file sizes which might be typical of PDF files. The highest bandwidth (10000 kbps) is typical of copying files over a local network. 1000 kbps would represent a decent broadband connection. 10kbps would represent a poor dial up connection and 1 kbps would represent a heavily shared connection such as at the Malawi College of Medicine. Note that these figures do not take account of protocol overhead, variations in available bandwidth, or interrupted connections, all of which will increase the required download time.

Bandwidth (kbps)	*Time for 100 kB*	*Time for 400kB*	*Time for 1500 kB*	*Bandwidth Equivalent*
10000	0.08 s	0.32 s	1.2 s	Local Area Network
1000	0.8 s	3.2 s	12 s	Broadband connection in developed world
100	8 s	32 s	2 minutes	ISDN connection
10	1 minute 20 s	5 minutes 20 s	20 minutes	Slow dial up link
1	13 minutes 20 s	53 minutes 20 s	3 hours 20 minutes	Heavily shared connection

As shown in the above table, for users on the slowest connection, a 1500 kB file would take at least 3 hours and 20 minutes to download. Even a 100 kB file would take nearly quarter of an hour to download.

Using the Information

The user may not be able to view downloaded PDF documents, since their computers may not have the necessary software installed. Obtaining a free viewer from the Internet requires another large download, assuming that the user identifies the problem and can find a viewer.

The user may not be able to use the text or images in another document. If the PDF document contains scanned images of the original document rather than plain text, it is not possible to access the text in the document without further processing. This requires Optical Character Recognition (OCR) software that is unlikely to be available to the user.

Text copied from a PDF document is often badly formatted, and images copied using the standard software (Adobe Reader) are of a lower quality than the original. Some authors use PDF document features to prevent users from extracting text or images, or printing the document.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Solutions

This section examines potential solutions to some of the issues discussed above.

Network Infrastructure and Support

A lack of bandwidth and reliable connectivity is at the core of most impediments to access. Dealing with these issues effectively would often involve long-term capacity building at a regional and national level, as well as simply upgrading the infrastructure within individual institutions, and as such is beyond the scope of this document.

However, even if more bandwidth were somehow made available to institutions, the situation would not necessarily be much improved. Our experience, and that of partner institutions, suggests that the bandwidth and connections that currently exist are often used inefficiently in a number of ways, to the point that improving network management would be far more cost-effective than any moderate increase in the maximum capacity of institutions' networks.

To achieve this improvement requires trained staff in combination with software and hardware to enable network management. Although the technical solutions must in the end be implemented at the user institutions themselves, this effort can be greatly assisted by the provision of documentation, tools, and other support such as on-line assistance and other outsourced services.

A longer term project looking into the provision of appropriate support for network administrators in developing world institutions could have great impact on the quality of Internet access. In addition, such a project would have other effects such as making the use of internal networks more efficient and practicable, and supporting the ongoing development of local technically skilled communities.

There are a number of solutions that could be put in place by institutions to prevent their available connectivity being wasted. These include:

- bandwidth quotas to prevent inefficient and unequal allocation of bandwidth between users
- anti-virus software and spyware scanning tools to detect and remove malware
- usage policies backed up by firewalls, to prevent unauthorised use of resources such as file sharing programs.

Effort to reduce the reliance on web mail could also result in significant bandwidth savings. This could be achieved by staff training and the provision of appropriate software. Alternatively, if the level of local technical support were not capable of reliably maintaining a local email system, off-site hosting of email services could be offered.

Although addressing the barriers to access within users' institutions, rather than on websites themselves, could require far more effort to do in great depth, those barriers should not be ignored. Even a moderate investment in addressing this situation, for example on-line documentation or links to download managers, could make a significant impact on the effectiveness of a website, and could have the long term effect of improving access generally, rather than purely to one specific site or service.

Site Navigation

Websites should be as easy to navigate as possible, taking into consideration compatibility with a variety of browsers and settings, in addition to bandwidth limitations. Commonly used areas of the site should be accessible within a single click from the front page, and navigation to any area of the site should be both intuitive and achievable in as few clicks as possible, to minimise downloading time.

In cases where websites link to documents on other websites, it should still be possible to access them in the minimum number of clicks and as intuitively as possible. Ideally, the document would download directly from a link on the first website, and not require the user to navigate the other website as well.

There are a number of guidelines for websites that impact on compatibility and bandwidth requirements, including:

- The web server should have compression enabled, so as to reduce the time it takes to download each page. All commonly used web servers support this, but it is not enabled by default.
- Images and objects should only be used where absolutely necessary.
- Alternative text should always be provided for accessibility purposes and for users browsing without images.
- Where large images are necessary, the user should be warned about the size of the page.
- Size information should also be given for any downloads.
- Javascript should not be required to access any information from the site.
- Using strict HTML 4.0 and a separate CSS file for layout further reduces bandwidth requirements, as well as increasing compatibility.

Websites should be tested with various browsers, including open source ones, to ensure that they are compatible with older computers and those running free software. Flash animations, audio and video should not be used. Where their use is unavoidable, a plain text alternative must be provided for users who cannot access them. The site could also be designed and tested to work well with the Loband service (www.loband.org), which simplifies web pages to reduce download times. This service could then be used to provide a text only version of the site without additional effort.

The site should include documentation or training materials to explain which resources are available and how to find them, including the use of any advanced search functionality. Documentation should take into account the differing levels of both computer literacy and literacy of potential users, particularly literacy in reading English if that is to be the only language in which a website is offered. Depending on the target audience, it may be necessary to make the website available in several languages.

Searching

A search box for the site should be visible from all pages, and particularly visible on the main page. There should also be a link to an advanced search page to allow users to perform more powerful searches to find documents more quickly and accurately.

The search result page should not contain so many results that it takes a long time to download. This could be made customisable by the user to suit their individual requirements.

Search results should contain a useful summary of the document in the form of a preview or extract in order to allow the user to select the most relevant documents to download. Because the PDF format forces the user to download images and formatting information, a text only or HTML version of the entire document may also be useful to the user as a preview or alternative. Both would offer reduced file size, and HTML gives the user the option of removing images if desired. As mentioned above, the size of every downloadable file should be made clear to the user.

Use of document metadata can also help users locate documents relevant to their needs. Document metadata is data that describes a document, e.g. subject, keywords, intended audience. The accuracy of the metadata is of key importance if this approach is used, particularly in the case of scanned documents, where no plain text is available for searching.

If this metadata were made publicly available in a machine-readable format, including the URL for downloading the document, this would make it possible for users to create new tools to find relevant documents, such as off-line search tools. This would also allow the documents to be listed by open archives, and shows support for such initiatives that reduce the cost of access to documents.

It would be possible to make an off-line search interface to a website or document repository, which installs on the user's computer, and allows them to search for documents without needing Internet access. This could be made available for download from the website, and could also be distributed on CD-ROMs.

Moving the on-line searching process to the local computer is much more efficient in situations where Internet connectivity is a problem, and where the site is likely to be used regularly within an institution.

Since a significant effort may be required to achieve this, it is often worthwhile to carry out research into patterns of usage by institutions currently accessing website content, to identify the number of regular users in low-bandwidth environments who might benefit from the off-line approach.

One problem with many off-line tools is that they become outdated and must be frequently replaced. Integration with an on-line repository can allow the tool to be updated on-line occasionally, in a way that does not require user interaction and therefore does not frustrate the user, and can also be more bandwidth-efficient. There is a tradeoff between downloading full copies of new reports, which requires more bandwidth, against downloading only summaries and metadata, which would still require the user to go on-line to download the full report if they decide that they need to read it.

Downloading

Download Management

To mitigate the problems of unreliable connections and time-dependent bandwidth availability, it is possible to use software applications called download managers. These client-side applications can queue downloads for greater efficiency; pause and resume downloads; recover interrupted downloads without restarting; and schedule downloads, thereby optimising bandwidth usage and reducing costs by downloading during off-peak hours. They often have the ability to integrate with web browsers. Some user participation is required to most effectively use download managers.

Many download managers already exist; a list is available on the Wikipedia website:

- http://en.wikipedia.org/wiki/List_of_download_managers

Most download managers can be downloaded and used without cost (free), although some contain advertising or spyware which is especially undesirable on a low bandwidth connection. It would be useful to develop a full set of requirements for download management software, and to carry out a full evaluation of existing software. If necessary, it would be possible to develop customised download management software.

Many users are not aware of the existence of download managers, or don't know which ones work well or are safe to use. Provision of appropriate tools and instructions on websites could lead to greater uptake, as would encouraging institutions to provide download managers and user education locally. The latter would reduce the bandwidth cost of each user downloading tools individually.

To increase the effectiveness of download management, it would be necessary to enable support for “HTTP Resuming” on the servers that store the PDF documents for download. This

allows download managers to resume an interrupted download, rather than restarting the download, which wastes the information already downloaded.

Some users dislike download managers and would prefer to have the document e-mailed to them. Indeed, it may be possible to use email as a method of transferring PDFs to the user more reliably than through an HTTP download, especially where the email server is physically close to the user, or the user's connection to the server is much better than their Internet connection.

The email server automatically handles the downloading of email on behalf of the user, and automatically retries a failed download for a certain amount of time before giving up. Once the email is received and stored on the local e-mail server, it could be much easier and faster for the user to download onto their local machine.

The request for the email containing the PDF could originate through the website, or through a locally installed search interface as described above.

This option has a number of disadvantages compared to a download manager:

1. Files sent as attachments to email require 30-40% more bandwidth to transfer, which will increase the time, bandwidth load and potential cost for the user;
2. Although the email server would automatically retry to download the file if the connection failed, it would start over from scratch every time rather than picking up where it left off, thereby again adding to the time taken, bandwidth overheads and potential cost;
3. Users would require sufficient space in their mailbox for the file, with the 30-40% overhead. If less space is available, then the download will silently fail;
4. This solution depends on the existence of local email servers, and their regular use by members of an institution.

Our experience is that in fact in many institutions, users prefer to use web mail because they do not perceive the local service to be reliable or trustworthy. There is no benefit whatsoever to using email delivery of documents to a web mail service. It would be important to ensure that this was understood by users if email were ever offered as a means of accessing stored documents.

Caching

In institutions or networks with many users, the same web page or document will often be downloaded several times by different users. Software known as a proxy cache can help to reduce bandwidth use, by keeping a copy of downloaded files for some time afterwards, and sending this copy to users who request a file that is already in the cache. This technology can assist the downloading of documents both directly by reducing the bandwidth required to download multiple copies of a document, and more generally by making the overall bandwidth usage of an institution more efficient.

Several proxy caches exist, and some are freely available. For example, some institutions may have purchased Microsoft Small Business Server, which includes their ISA server proxy cache. The open source software Squid is available for Windows and most Unix servers, free of charge.

In most cases, to take advantage of a proxy cache, each user's computer will have to be reconfigured to send all requests for web pages through the cache. The only alternatives are to use Internet Explorer's autoconfiguration to allow it to detect proxies automatically, or using the transparent proxy features of some gateway software. If this is not done, then most users will ignore the cache and access the documents directly, eliminating any potential bandwidth saving.

In order to work properly, a proxy cache must be able to identify when a user requests a document that is already contained in the cache. This can be defeated by websites that vary the address of documents by including a unique identifier that varies over time. Every effort should be made to avoid using variable (dynamic) addresses to refer to documents on websites.

It is also possible for caching to be defeated by firewalls (network security devices) on the server or client side that modify the request. We are not aware of any existing software to test that caching is working correctly, but it can be tested manually, and software or instructions could be developed.

Some proxy cache software allows the administrator to cache certain types of documents for longer or shorter periods, which might help to keep PDF documents in the cache for as long as possible. We are not aware of a freely available proxy cache which can do this, but it would be possible to customise an existing open source cache to give the user additional control.

Where a set of documents has consistent features, such as the same embedded fonts and images, it would be possible to accelerate downloads of subsequent documents by reusing the portions which are duplicated between files. This could be done using some open source software like rsync, but to obtain maximum efficiency and ease of use, it would be necessary to develop custom software that understands the structure of PDF documents.

There is some potential for integration or overlap between download management and proxy caching, since a cache (or Library Management System) must also download the document at least once. There is also the potential for conflict, as some download managers run on end-user computers may not be able to take advantage of, or work correctly with, proxy caches. One way to take advantage of both technologies would be to integrate the download manager functionality with the proxy cache. In this situation, the user might retain scheduling control, i.e. *when* files were downloaded, e.g. for off-peak optimisation, while the proxy cache handled the actual transfer of the files, using local copies where possible, and resuming downloads as necessary.

Incorporating download management with an off-line search interface that covered both on-line content and files already available locally would be another means of combining what is in effect a long-term and searchable cache along with efficient downloading. Looking to the more advanced end of the spectrum, a fully integrated Library Management System such as ELIN, developed by the University of Lund, or similar could be used. This could incorporate advanced search, download management, long-term local storage of documents institution-wide, along with other library-specific functionality.

It would be useful to produce a set of requirements for caching systems, and to evaluate existing proxy cache and library management software accordingly. This could form the basis of appropriate recommendations to institutions using websites that could benefit from them. Software that meets the identified requirements could be made available for download on the website, or distributed to users on CD-ROMs.

Implementing caching, download management or integrated library systems, as with any solution that requires uptake by the user institution, would depend on a level of available technical support and resources to enable user education. It is important that this is evaluated when considering proposed solutions to identified user requirements.

Reducing File Sizes

Users will have the majority of problems downloading large document files. These are the most likely to be interrupted during the download, and will take the longest time to download over a slow connection. Therefore, it will help if the size of documents is reduced as much as possible.

Large PDF documents are particularly problematic because the user must download the entire document to access any part of it, as opposed to HTML where users can choose to avoid

downloading images. PDF documents can include additional information such as embedded fonts and colour profiles, which help to ensure the highest quality of display and printing. These features consume additional space in the file and therefore must be downloaded by the user.

It is possible to address issues of file size at both the point of creation and on the server where the PDF is to be downloaded.

One method of reducing PDF document sizes would be to produce guidelines for authors who create these documents. The guidelines could advise authors on methods to avoid creating files that are larger than necessary. For example, they might avoid using optional elements; use standard fonts; avoid repeated images; and create the documents using high quality software such as Adobe Acrobat, which is very good at minimising the file size.

Images can constitute a significant proportion of the total size of a PDF document. Images can be compressed at different levels to trade off image quality against file size. Different users will have different requirements for image quality and file size, and it would be preferable to offer a range of downloads. It is possible to configure Acrobat to produce documents of different levels of compression and quality from the same original content. We are currently unaware of any automated tools for increasing compression of existing PDF documents, however these could be developed. Alternatively, the guidelines could require authors to produce a number of documents with different compression levels.

It is possible to split large PDF documents into several files, which can be useful if the content is clearly delineated, and where some sections are useful independently of others. An obvious example is to offer the main content of a document separately to the appendices. Software tools exist to save a portion of a PDF document separately. Due to the subjective nature of this modification, it is probably best done by the original author. Again, the guidelines could advise authors when to do this.

A tool to analyse the use of space within a PDF document may be of use to authors. This tool would complement the given guidelines by detecting the presence of unnecessary options or unusually large images. Users of Acrobat Professional can use the built in space auditing tool to achieve this manually. We are unaware of any tool that offers suggestions to PDF creators to make PDF documents fit specified guidelines; it would be possible to create such a tool.

Another way to reduce file size is to offer the text of the document rendered as HTML or plain text as smaller alternatives to the PDF documents. It is easier to produce high quality HTML documents directly from the original content, and it is worth considering HTML as an alternative publication format to PDF when creating a document: a good summary can be found on this website:

- http://www.alistapart.com/articles/pdf_accessibility/

Consideration of the appropriateness for different file formats when presenting information could form part of the guidelines to authors. PDF documents can also be automatically converted into HTML using software tools, with some loss of quality. Examples of such software already exists, and Adobe runs a web service that performs the same functions, although this service is not free. It would be possible to evaluate existing tools for use with the material on a website, and to create new tools or online services if required.

As well as reducing download time, providing text or HTML versions of PDFs could improve compatibility with accessibility technology. For example, some screen readers may work better with text or HTML content than with PDF documents.

Alternatives to Downloading

It is possible to avoid many of the problems associated with bandwidth issues and download requirements by distributing the contents of a website on CD-ROM discs. PDF documents could

be distributed along with the off-line search functionality mentioned above, with the ability to integrate with the on-line website to allow a search to return results from both off-line and new, on-line documents. Useful software such as PDF document viewers, library management and download management and caching functionality could also be provided on the CD-ROMs, as could documentation and educational materials. The CD-ROM set could be made available through the website, in addition to on-line, downloadable versions of the tools and documentation.

While Digital Versatile Discs (DVD) has a higher capacity than CD-ROMs, DVD drives are not commonly found in developing countries, and the discs are much harder to copy for backup or distribution within an institution, so we would recommend against using DVD distribution at this time.

Using the Information

Making PDF document viewing software available for download from a website would be useful for users of that site who do not currently have such software installed on their machine. Different PDF viewers have different feature sets and different download sizes; the latest version of Adobe Reader is probably the largest example. It would be useful to identify requirements for PDF viewers and to offer a selection that reflects user needs.

As different PDF viewers support different versions of the PDF standard, it would be useful to encourage authors to generate documents compatible with older versions. There is little reason to require viewers compatible with versions of the standard later than 1.4 since the additional functionality is unlikely to be necessary, and will result in larger documents if used. Most installed PDF viewers are likely to be compatible with at least version 1.4, meaning that no further download will be required. In addition, viewers that support the latest versions of the standard are likely to be larger and will therefore be more difficult to download.

Furthermore, the PDF-Archive standard currently being developed is intended to be an efficient format for the long-term storage of documents, and will consist of a reduced subset of the 1.4 standard. Therefore, requiring viewers compatible with 1.4 or less would be compatible with a possible shift toward using PDF-Archive as a future standard.

Guidelines to authors could require documents to be free from protection against copying text or printing, in order to ensure that the text can be easily used and accurately quoted in future documents.

The standard technique to allow access to text in a scanned PDF document is to convert the images to text using Optical Character Recognition (OCR). The text is then made part of the PDF document as metadata. This can be done either manually, for greater accuracy, or automatically, for lower cost. This allows for full text searching, potentially improving search accuracy. Additionally the text could then be made available separately to the document as alternative formats such as text or HTML, to improve access time as detailed previously.

Recommendations

In summary, our main recommendations are:

- **User Engagement** - identify a group of users and work with them to get ongoing feedback;
- **Website Optimisation** - ensure that website interface and content is as accessible as possible, particularly over low bandwidth connections;
- **Local Infrastructure Support** - provide support to users and user organisations in optimising their network usage; and
- **PDF Document Optimisation** - work with content providers to ensure that PDF documents are appropriate in terms of size and metadata.

User Engagement

Working with users is vital to ensure that the outcomes of any project are actually useful for users. We believe that the best way to manage projects is to work iteratively and evolve solutions. Each version should be put to use by the user group and their feedback should inform the priorities for features for the next iteration. Regular input from a variety of users will ensure that project work will remain focussed on the key barriers to access. This point underpins any proposed solution.

Website Optimisation

Our philosophy is that websites should be simple, in terms of the page content, the navigation, and the structure of the site. This decreases the amount of data downloaded for a user to use the site, and makes it more accessible to people with disabilities, as well as those on low bandwidth. The site will be more compatible with old software and less powerful computers. Search functionality will reduce the amount of time spent on-line browsing documents, while multiple versions of documents will optimise the on-line experience for users with different bandwidth requirements.

Techniques to ensure website accessibility include:

- enabling compression on the web servers;
- reducing the size of pages;
- making it easy and intuitive to navigate to the required content;
- powerful search capability with results including extracts from the documents;
- providing education material on using the site and complementary tools.
- providing metadata and useful summaries of documents;
- making extracts of documents, or whole documents, available as text or HTML;
- making text content available for scanned documents;
- using tools to further compress existing documents.

Local Infrastructure Support

While issues purely related to the design of websites are probably most easily addressed, it should be recognised that most of the limitations are on the client side of the network, and that unless similar attention is paid to this side of the issue, only a limited increase in accessibility will be possible.

From previous fieldwork we have found that the lack of network administration skills and infrastructure in the developing world has a huge impact on the bandwidth available for end users. All projects aimed at improving access from the developing world should consider building capacity in the area of network skills and infrastructure.

We would expect useful solutions for user institutions to include the following:

- Create a local search tool so that users can perform searches off-line.
- Identify or create download management software optimised for website users.
- Integrate download management and local search tools.

PDF Document Optimisation

The fourth key area for potential improvement is analysis of existing material for accessibility, and where necessary, working with content providers to ensure that PDF documents produced are accessible. Guidelines for authors should ensure:

- documents optimised for size;
- meaningful summaries and metadata;
- where appropriate, multiple versions with different levels of image compression, or even no images;
- HTML versions where appropriate, either instead of or in addition to PDFs.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Summary

Websites intending to distribute information to the developing world face a significant challenge, because many users in these countries will have poor connections, poor infrastructure and/or poor network administration.

The documents to be disseminated are often in Adobe Portable Document Format (PDF). These files tend to be large, and can be difficult to access for users with slow and intermittent connections.

There may be ground to be gained in producing guidelines to standardise content and support document authors. However, many PDF documents are of reasonable size for the information contained. It may be possible to remove some information, such as images and fonts, to make downloading easier.

It seems useful to widen the scope of potential improvements to include the design of such websites themselves, and what can be done to support users in gaining access to the documents.

We recommend that organisations operating such websites should investigate the bottlenecks in accessing documents for their target audience, and the local environment in which solutions can be provided. This will involve working closely with users to understand their requirements.

We recommend that:

- compression is enabled on your web server(s). This is a very quick task that could double the speed of access for many users;
- a group of users should be found who can provide ongoing feedback;
- websites should be optimised for users with low bandwidth;
- client side tools are provided to support downloading, managing and searching for PDF documents;
- support for authors is provided so that they will create appropriately optimised content.

References

Information About the PDF Standard

http://partners.adobe.com/public/developer/pdf/index_reference.html
PDF Reference documents

http://en.wikipedia.org/wiki/Pdf
Wikipedia page about PDF

http://www.iso.org/iso/en/commcentre/pressreleases/2005/Ref974.html
Press release about PDF/A standard. Source of "the surface Web is 167 terabytes ... 9,2 % of which consist of PDF documents."

http://www.pdfzone.com/article2/0,1895,1885626,00.asp
Advice about how to reduce PDF file size.

PDF Tools

http://en.wikipedia.org/wiki/List_of_PDF_software
List of PDF software from Wikipedia

http://createpdf.adobe.com/
Adobe "Create PDF" service (from various office formats, DTP, CAD, PostScript, image ...)

http://www.adobe.com/products/acrobat/access_onlinetools.html
Adobe "Convert PDF" service (to HTML or text).

http://www.apagoinc.com/PDFEnhancer
A tool that will reduce the size of PDF documents.

Download Managers and Caching

http://en.wikipedia.org/wiki/Download_manager
Wikipedia page about download managers.

http://en.wikipedia.org/wiki/List_of_download_managers
Wikipedia list of download management software.

http://www.squid-cache.org/
Squid is an open source web proxy cache for Linux and Windows.

http://www.lub.lu.se/headoffice/elininfo.shtml
The ELIN@ service "satisfies the end-user demand for one entry point to federated searching across multiple digital resources".

Metadata

http://dublincore.org/
Dublin Core is the main metadata standard for documents.

http://en.wikipedia.org/wiki/Dublin_Core
Wikipedia page about Dublin Core.

http://www.openarchives.org/
The Open Archive Initiative maintains a protocol for metadata harvesting.

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org

Universities and Information Access in Developing Countries

http://www.inasp.info/pubs/bandwidth/index.html
“Optimising Bandwidth in Developing Countries' Higher Education”, INASP

http://www.inasp.info/psi/ejp/index.html
“Electronic Journal Publishing Reader”, INASP

http://www.inasp.info/pubs/INASPdigitallib.pdf
“Towards the digital library: findings of an investigation to establish the current status of university libraries in Africa”, INASP

Access Issues

http://www.alistapart.com/articles/pdf_accessibility/
This article reviews PDF documents from the point of view of disability access. It states "Most PDFs on the web should be HTML." .

www.aidworld.org The Coach House, 6 Chaucer Road, Cambridge, CB2 2EB, UK +44(0)1223 331716 info@aidworld.org