Scoping Document on PDF Optimisation

This document explores potential issues that users in the developing world might face in using DFID's Research Portal, particularly regarding the downloading and usage of research papers as PDF documents.

Table of Contents

Glossary

Term	*Definition*
AGORA	Access to Global Online Research in Agriculture
CABI	CAB International
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
DFID	Department for International Development
FAO	UN Food and Agriculture Organisation
Firewall	A piece of hardware or software that prevents unauthorised connections being made in a network environment.
HINARI	Health InterNetwork Access to Research Initiative
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.
INASP	International Network for the Availability of Scientific Publications
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
NGO	Non Governmental Organisation
OCR	Optical Character Recognition
PDF	Portable Document Format
PERI	Pharmaceutical Education and Research Institute
Proxy	A network service allowing computers to connect to another service indirectly.
UN	United Nations
XML	eXtensible Markup Language

Introduction

"*The inability to download large PDF files from publisher web sites is the greatest constraint to the effective use of AGORA after the availability of reliable and good quality Internet access, noting that the two are clearly interrelated.*" - Stephen Rudgard, UN FAO.

This document explores potential issues that users in the developing world might face in using DFID's new Research Portal, particularly regarding the downloading and usage of research papers as PDF documents.

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 on the portal 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 through the Research Portal. 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 the portal. These include:

- Work closely with users to identify their expectations and intended use of the portal, and any problems they experience;
- Ensure that development of the Research Portal and related tools is driven by user needs;
- Provide client-side tools to improve the efficiency and reliability of searching for and downloading files;
- Ensure that information presented through the Research Portal, and the portal itself, is 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 the knowledge resources soon to be available through DFID's Research Portal.

Existing projects that have created research portals 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 useable, thereby reducing the effectiveness of these projects. In order to meet its objective, this document will identify methods of overcoming these factors. To achieve this, the scope of this report is wider than the original terms of reference. The terms of reference refer to:

- methods of producing smaller PDF documents and
- download and bandwidth management.

This document also covers

- the Research Portal interface;
- the search interface to material contained on the Research Portal;
- the requirements of users viewing and working with PDF documents after download.

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 online 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 online content as PDF, particularly for documents that are primarily formatted for printing. In fact, most documents stored online 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.

As an example of the format's popularity, 48 of the first 50 documents accessible from the publications section of the dfid.gov.uk site are PDF files. 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 criticised, often 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 online documents.

Access to Knowledge Sources from Developing World Institutions

Several initiatives aim to enable institutions in developing countries to access knowledge resources such as research documents. Examples include the PERI, HINARI and AGORA projects, the INASP African Journals Online project, and the forthcoming DFID Research Portal. All of these provide access to information in electronic form over the Internet.

In order to address the aims of poverty reduction and information transfer, it is important to cater to the needs of users and potential users working in the low bandwidth and intermittent connectivity environments often found in institutions in the developing world. It is therefore essential that the process of accessing the information contained in PDF documents through the Research Portal is made as easy as possible, with the needs of developing world users in mind.

There are many potential usage scenarios for online PDF documentation. It is important to consider the different current and potential user and institution profiles when addressing access issues. Potential users of the DFID research portal include students, researchers, NGO workers, DFID 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.

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.

Issues

Users in the developing world often have to overcome a number of constraints to access information on the Internet. Reaching these users represents the greatest challenge for the DFID Research Portal.

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 the Research Portal.

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 webmail 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 webmail 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 of the Research Portal may have basic standards of computer literacy, 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 download time against bandwidth for files that would be considered small or moderately sized in the context of the Research Portal material. The highest bandwidth 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. The real time is likely to be longer.

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 can be seen, for users on the slowest connection, a 1500 kB file would take 3 hours and 20 minutes. 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.

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 online 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
- antivirus 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 webmail 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 the Research Portal could require far more effort to do in great depth, these should not be ignored by this project or others like it. Even moderate or small investment in research and solutions addressing this situation, for example online guidance, could make a significant impact on the effectiveness of the portal, and could have the long term effect of improving access generally, rather than purely to one specific site or service.

Site Navigation

The website used as the interface to the Research Portal should be as easy to use 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 the documents are available on third party websites, for example articles contained in documents on publisher's websites, it should still be possible to access these in the minimum number of clicks and as intuitively as possible. Ideally, the article would be directly available for download from the Research Portal site, rather than the user being redirected to another website.

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

- The 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 the 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 offline 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 research.

It would be possible to make an offline search interface to the Research Portal. This could be made available for download from the website, and could also be distributed on CD-ROMs. This would move the online interaction required to use a web interface to the local computer, which is much more efficient in situations where Internet connectivity is a problem, and where the interface is likely to be used regularly within an institution. Where institutions may use the portal infrequently, this approach may be less useful, as the time taken to install a local interface may prove as much of an obstacle to increasing usage as bandwidth issues. This approach would therefore be most useful to institutions that are likely to be regular users of the Research Portal. It would be useful to carry out research into patterns of usage by institutions currently accessing Research Portal content to evaluate the usefulness of this approach. The ability to interface to and search the Research Portal archives locally would reduce the time and bandwidth required to locate relevant documents. This search tool could be integrated with the online repository, meaning that the local database of content could be updated when the user was online, ensuring that the local content was kept up to date with the most recent uploads to the server. This would require the accurate and full provision of metadata to be most useful, and would not allow the full content of documents to be included in searches. However, through limiting the amount of interaction with the repository that must happen online, the user would reduce the total amount of time spent accessing documents by a significant amount.

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 at http://en.wikipedia.org/wiki/List_of_download_managers . Most are available for 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 know which ones would be trustworthy or most useful. Provision of appropriate tools and instructions on the portal 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.

If an institution has a locally sited email server that centrally collects email, it may be possible to use email as a method of transferring PDFs to the user more reliably than through an HTTP download. If it were possible for the user to request a PDF be emailed to them, the email server would handle the downloading of the file, and it would then be available for the user to download onto their local machine from the local server. The request for the email containing the PDF could originate through the Research Portal, or through a locally installed search interface as described above.

This would be a less preferable option than a download manager for a number of reasons. Firstly, because of the way email handles attachments, sending files by email can increase the amount of data which must be transferred by 30-40%, which will increase the time, bandwidth load and potential cost for the user. Secondly, 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. A further point is that users would have to have a mailbox size allowance great enough to allow them to download the file - if this were not the case then the download would fail. Furthermore this solution would depend on the existence of local email servers collecting mail centrally, and being regularly used by members of an institution.

Our experience is that in fact in many institutions, users prefer to use webmail because they do not perceive the local service to be reliable. If users attempted to use a document-by-email service over webmail, they would significantly increase their overall bandwidth requirements without any benefit, as the connection to the webmail service would still be as prone to interruption, only with more data to download due to the email wrapper for the file, in addition to having to access the webmail site itself. It would be important to ensure that this information 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 addresses to refer to documents on the Research Portal.

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 DFID Research Portal 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 offline search interface that covered both online 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 would assist DFID in making appropriate recommendations to institutions using the Research Portal. Software that meets the identified requirements could be made available for download on the Research Portal 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 documents for the Research Portal. These 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 at 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 the Research Portal, 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 the Research Portal on CD-ROMs. PDF documents could be distributed along with the offline search functionality mentioned above, with the ability to integrate with the online portal to allow a search to return results from both offline and new, online 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 research portal, in addition to online, downloadable versions of the tools and documentation.

Using the Information

Making PDF document viewing software available through the Research Portal site would be useful for users that 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.

The standard technique to allow access to text in a scanned PDF document is to convert the images to text using 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;
- **Portal Optimisation** - ensure that the Research Portal 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.

Portal 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. It makes the site 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 online browsing documents, while multiple versions of documents will optimise the online experience for users with different bandwidth requirements.

Techniques to ensure portal 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 Research Portal are probably most easily addressed, it is likely that issues at the user end will have to be addressed in order to most effectively improve access to the portal. 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 offline.
- Identify or create download management software optimised for Research Portal users.
- Integrate download management and local search tools.

PDF Document Optimisation

The third 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.

First Steps / Project Plan

As an example of how our recommendations might be put into practise, we have drafted a project plan that takes into account the points raised.

Phase	***Task***	***Notes***
1	Identify current / potential user group(s) to work with	Identify representative groups based on local size, and national network infrastructure, technical support. Work with multiple groups if possible to better inform development of solutions.
	Send team to work in a user institution	E.g. a university in a developing country.
2	Improve the Research Portal	Start with low hanging fruit, such as enabling server compression, basic site usage instructions, and making web pages and navigation simpler. User feedback and experience of using the site from a user institutions feeds into the work.
	Develop tools on site	This will allow the team to see how implementing a tool, for example a cache, affects the user experience. It will also allow understanding of any issues that arise with installing tools on the local network and educating users in their use. It also allows for a tight feedback circuit with the users.
	Review existing PDF documents	Are they optimised for size?
3	Continue tool development off-site.	Still informed by regular user feedback from contacts made during initial visit. Consider further trips to other institutions to check that assumptions of relevance hold across different environments.
	Continue website improvement	Evaluate potential improvements to search interface. Make different versions of documents available, e.g. HTML, different compression rates. Update site usage information to cover new developments.
	Produce PDF guidelines	If found to be necessary during phase 2.
	Automatic PDF Enhancement Tools evaluation / development	If found to be necessary during phase 2, with feedback from authors.

User feedback must inform all stages of this work. This is the rationale for sending a team to work in user institutions. The team will have to deal with the same problems as users, and will be able to rapidly integrate user feedback into the development of solutions.

The plan should be flexible so that the user feedback can direct the outcomes. For example, if the review of existing PDF documents finds they are near optimal already then there will be no need to produce guidelines for authors or tools to enhance or evaluate existing documents. The metric of success for such a project should be considered to be the effect it has had on the overall goal, i.e. improvements to access, rather than specific predetermined outcomes that may turn out to be less useful means of addressing user needs. A measurement of success of this project might be the increase in downloads from target institutions, or the decrease in average time taken to download documents.

Summary

The goal of the proposed DFID Research Portal is to disseminate research carried out by or for DFID. The audience includes users in the developing world. This is challenging because many of these users will have poor connections, poor infrastructure and/or poor network administration.

The research to be disseminated is generally in PDF form. PDF documents tend to be large, and can be difficult to access for users with slow and intermittent connections.

While there may be ground to be gained in standardising content and supporting authors, the examples of PDF documents from the portal that we have seen have been of reasonable size, so the problem will not be solved simply by making smaller PDF documents. Furthermore, we must not limit the scope of our thinking to the Research Portal itself, but rather investigate what can be done to support the users in gaining access to the documents.

To support the goal of the Research Portal we need to better understand the bottlenecks in accessing documents, 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;
- the Research Portal 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.

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." .