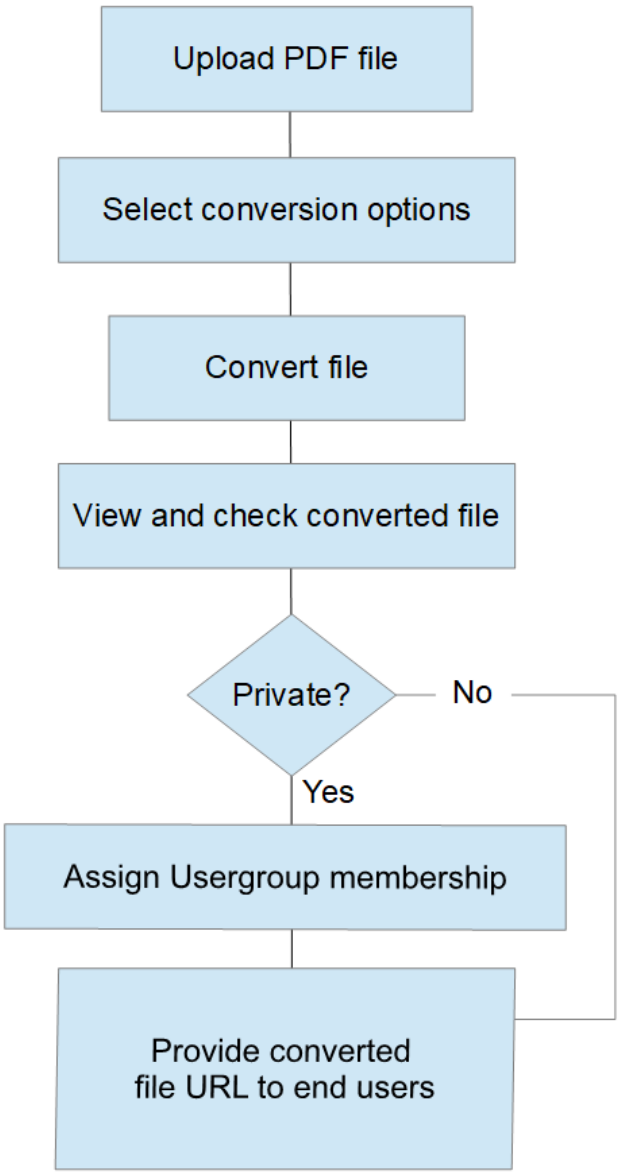
User Guide to Online Publishing with Webdoxx

Overview

Webdoxx PDF2HTML5 is an automated web-based publishing service that enables organizations and individuals to make their content available to either public or private audiences with a range of security protections. Typically, the source material will be industry-standard format PDF files, but other file formats such as PowerPoint, Word, ePUB etc. can be handled (on request) via our "managed Webdoxx" services.

This document provides a brief guide to the way Webdoxx PDF2HTML5 works and how to use it. Our Webdoxx managed services provide similar functionality, with a range of additional capabilities that are described later in this document. A visual summary of the main steps is shown below:



How it works

Source PDF files are the starting point for most publishers. These may be relatively short documents, like newsletters, brochures, legal or financial documents, journals and training materials, or much longer documents, such as textbooks and complex reference works.

The original creation of these PDFs will typically be a desktop publishing system like InDesign or Quark Express, or an office document creation system, such as Word, PowerPoint or OpenOffice. Ideally the PDFs that are generated from these source editors will be produced in a manner that recognizes that they will be used "on screen" rather than in print form and are optimized for this usage when generated. To this end it is recommended that generated PDFs conform to the recommendations in Appendix 1 of this document. These are recommendations, not essential requirements, so do please contact us if you have any questions or issues with your source material.

Webdoxx services work by pre-converting the source PDF pages to a matching, highly structured, HTML5 page format. Associated content, such as explicit internal and external links, annotations, bookmarks and all the text, are analyzed during the conversion process and stored in separate files. The resulting fileset can then be displayed in the Webdoxx secure document viewer. This provides near-instant response with very high-quality display, even when zooming in on text and images. However, for security and copyright reasons the source PDF is NOT retained – in our automated service the uploaded PDF is deleted immediately after conversion is completed, so there is no risk of the source file being accessible to anyone else (including our own team). *The Webdoxx viewer by default does not permit copying, printing or downloading, and for private files will prompt for username/password details (which are then validated for the document in question) before displaying content to the user. Variation of these defaults can be provided on our "managed Webdoxx" services, for example to facilitate accessibility by screen readers.*

For our managed Webdoxx services, publishers often prefer to send us their PDFs for processing by our specialists, after which we upload them to the target server, again without including the source PDF.

Once a converted fileset has been created on the service, it is available for display to end users by providing them with a web link (URL) to the converted document. There is no special software required, no plugins, and no need for special technical support:

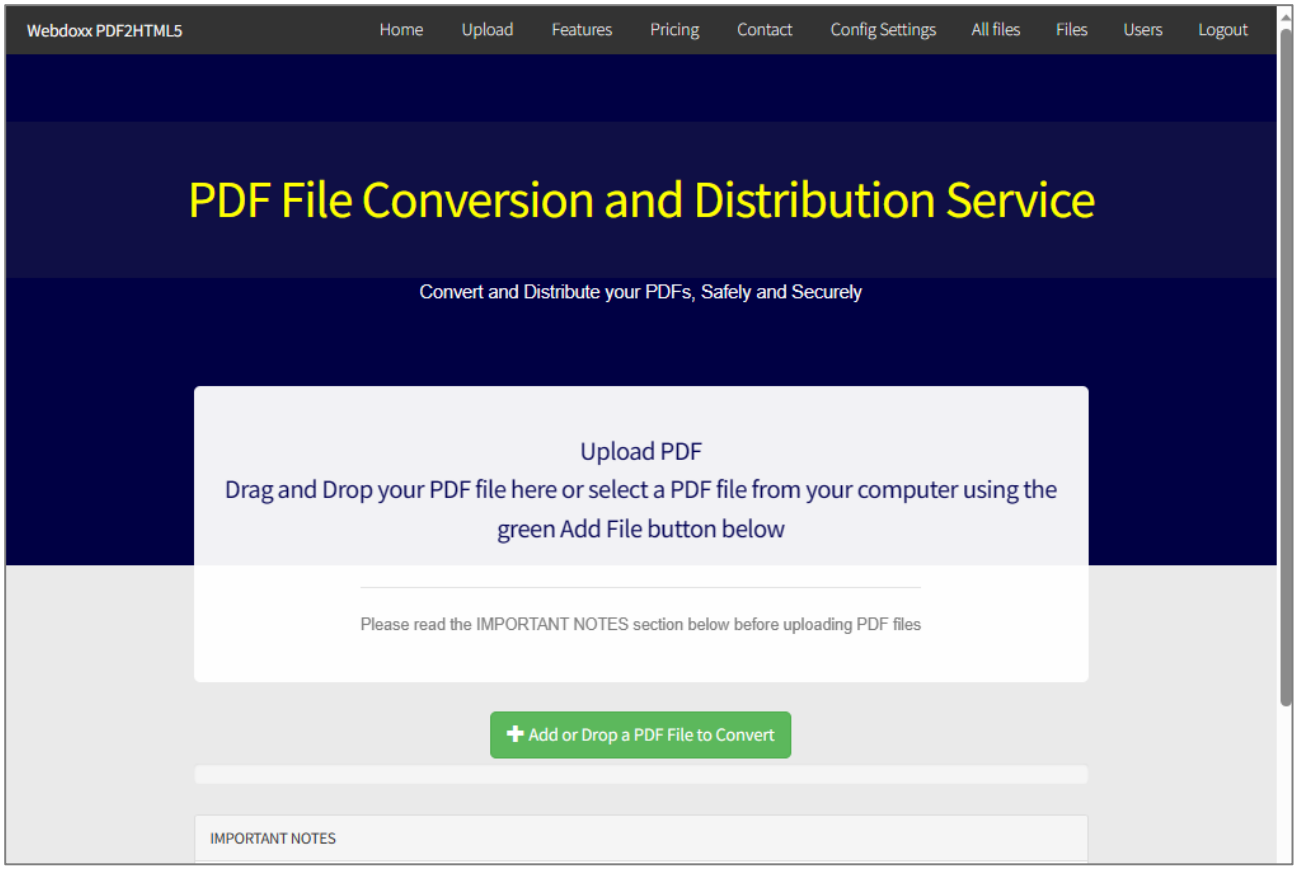
- If the file has been converted as a PUBLIC document, then anyone with that link can view the file using a standard HTML5-compliant web browser – in practice this means all web browsers on all technology platforms
- If the file has been converted as a PRIVATE file, then only end users who are registered on the appropriate Webdoxx service with permission to view that document will be able to access it, and that access is monitored and logged by the service. Private access files can only be created by registered users who are logged-in to the service

The various steps in the graphic at the start of this document are now explained in more detail.

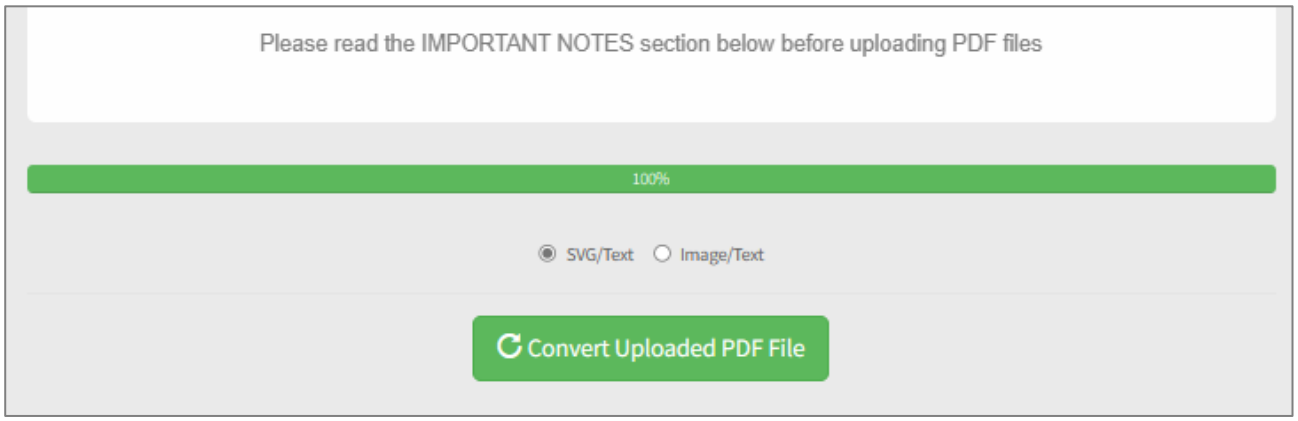
Getting started

The basic steps to convert a PDF ready for use are as follows:

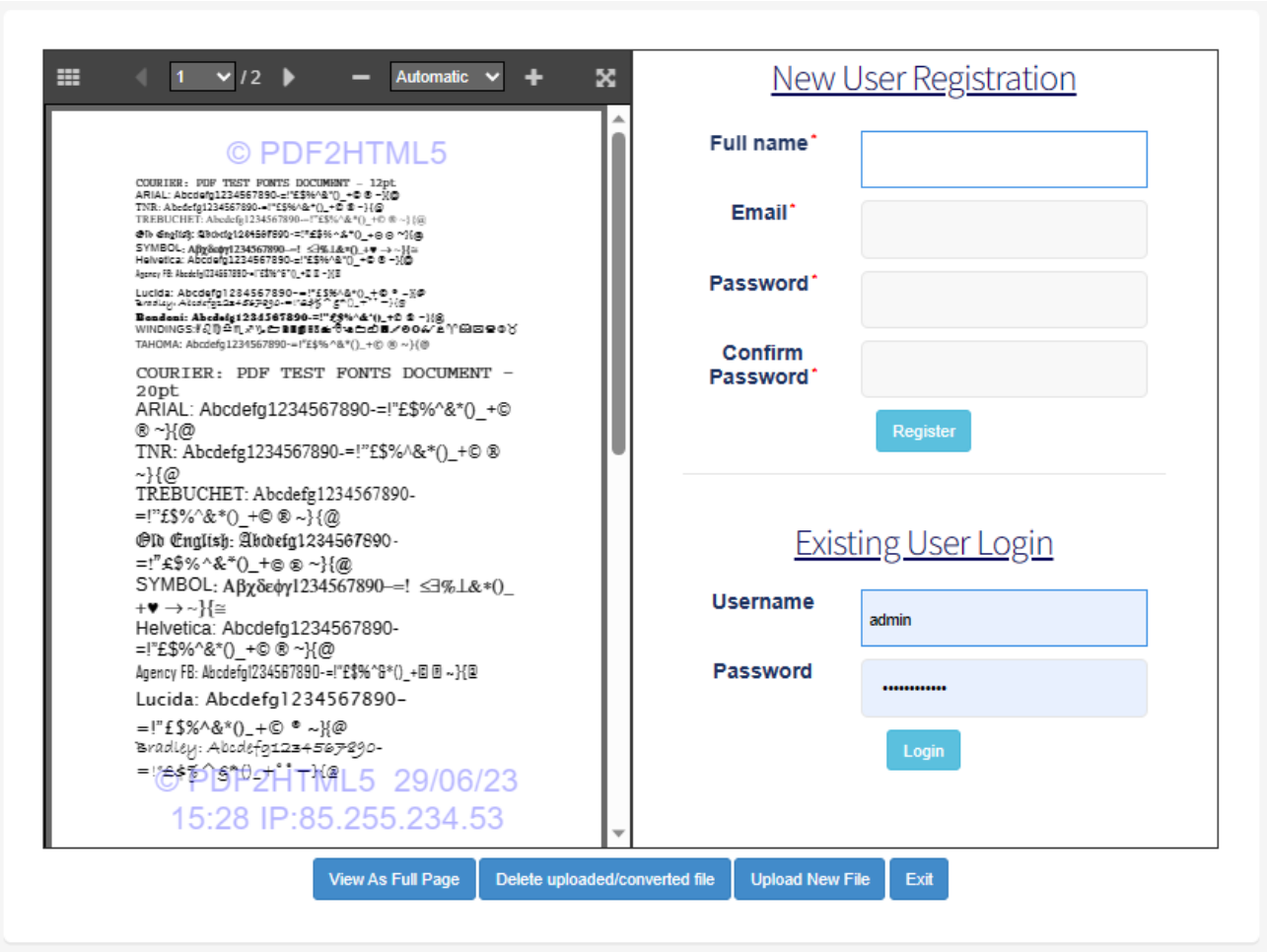
1. **Prepare the source PDF**: Make sure the source PDF has a convenient name - preferably not too long, and ideally without any special characters or spaces in it – the service will try and normalize the name of the file if necessary; and also make sure the source PDF is not set with Adobe-style security, and meets the length and filesize specifications for uploading on the automated service (for most files this means under 64Mbytes and less than 1000 pages). Long documents take longer to convert than short documents, so it is best to start/test using sample short documents. Free services are restricted to smaller files, and much larger files can be supported, but must either use our managed Webdoxx service or be reduced in filesize first – contact us if you need assistance and refer to the recommendations in Appendix 1 of this document
2. **New (unregistered) users:** Select the "Free & Subscription Services: Upload & Secure PDFs & Office files" link on the home page of the Webdoxx PDF2HTML5 service for direct access to the FREE test service. If this link is selected a "freeupload" page is provided (as illustrated below), and the PDF can be uploaded to our server. The IMPORTANT NOTES text on that page explain about file sizes, conversion options etc. Once uploaded the option to convert the file is provided, and the file will then be converted and displayed when complete.



After uploading a free/test file as described above, the option to convert the file is presented, as shown below:

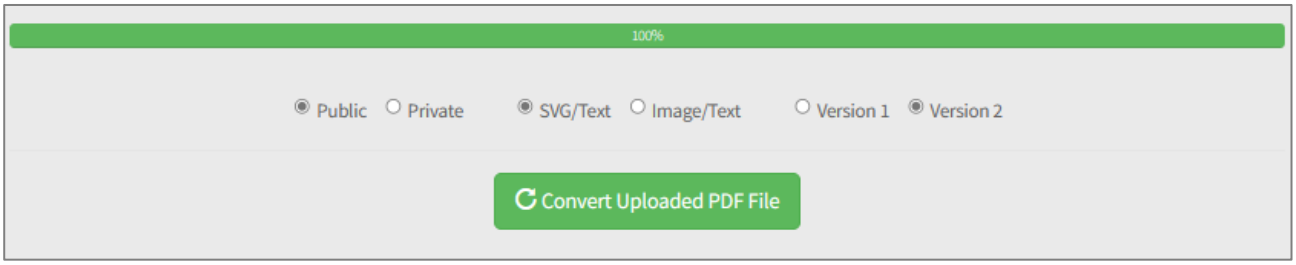


The default conversion mode (svg/text) is shown, which is suitable for almost all files, and when the green Convert button is pressed the file conversion commences. After a short period, the conversion is completed and a screen similar to that shown below is displayed (for logged-in/registered publishers just the converted screen element is shown). The buttons at the foot of this screen provide the option to "View as full page", which is the usual/recommended choice as it will show you what the page will look like to end users and the Address bar of the browser will include the full URL that you may choose to copy for providing to end users and for your own reference:



Publishers who are not yet registered can opt at this point to provide a username/password and email details in order to register and receive trial usage for a longer period, i.e., for testing both public and private access. Free (test) files like this are only retained for a short period (up to 7 days) before being automatically deleted by the service – for production usage of the service, publishers need to register first and, in most cases, become a subscriber

3. **Registered users:** Registered users should LOGIN to the service before uploading files. Login can be by username or email address, and password. Logged-in users then select the UPLOAD file menu if necessary and the Upload page for logged-in publishers will be displayed (this shows the same initial content as that displayed above). However, after uploading a PDF file the conversion section provides additional options, as illustrated below:



   This time you can choose to create public or private documents (private documents require users to login to the service), and you can also opt for different conversion settings – in most cases the default settings will be fine. The file may then be viewed when converted, for example here (a Public access example with default settings). Registered users can see details of all the files they have created via the FILES menu, where files can be viewed and access managed. And because the service knows your email address, it will automatically send you an email with a series of links to the converted document, each corresponding to a different display format: the default format ("complete"), a simplified display ("clean"), a magazine page-flip display ("magazine") and a slide format display ("slide")

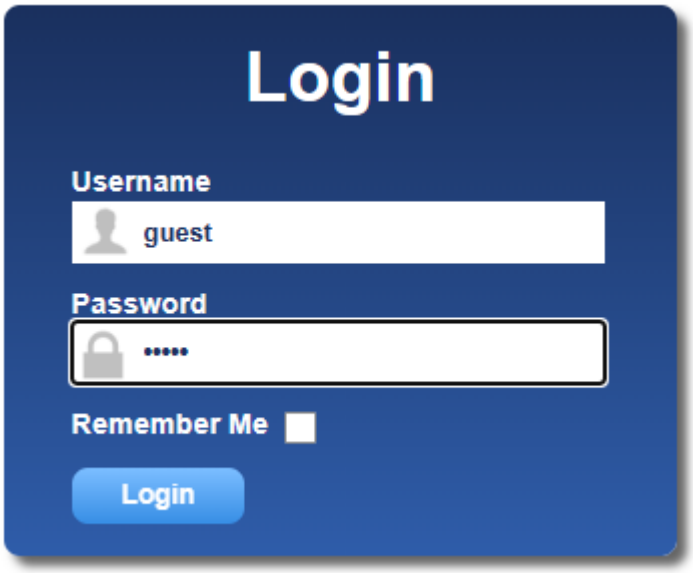
4. **Subscribers**: Registered users who are also subscribers to the service can create more documents, longer documents and control access to private documents. These facilities are provided using the Webdoxx PDF2HTML5 User Management and File Management facilities (see also, the sections on these topics below). The Webdoxx PDF2HTML5 service provides simplified User Management facilities, as does the default Webdoxx Managed service, but for customers requiring greater user management and file management control, dedicated (extended) facilities are provided. These are described in more detail later in this document.

Before sending a document access link to end users, please check the document thoroughly to make sure all pages have converted correctly, and any facilities that are included, like hyperlinks and bookmarks, are handled as you would expect. If there are any problems, first check the source PDF against our recommendations in Appendix 1, and if necessary, re-upload it after fixing any identified issues, and delete the previous version(s) via the FILE

menu facilities. If there are problems you cannot explain, just let us know and we will help resolve these. Each time a new document is uploaded on the automated service it will be assigned a new unique URL. Publishers who wish to keep the same URL or a specific URL for their documents, should use our managed Webdoxx services, as these provide for publisher-specified document storage arrangements, ensuring URLs match their particular requirements.

Login and autologin

Access to files that have been specified as Private require a user login. The default arrangement is that when a user selects the web address for the document (URL) a pop-up login form is shown, similar to that shown below (on our managed services bespoke login forms can be provided).



In general, the username field can be their registered email address or their registered username (if different). The password field will be an alphanumeric string provided when the user record was created (the service will generate an initial password if required). If the URL is something like:

https://www.pdf2html5.com/pdfupload/server/php/uploads/admin_qctcrbtsos/alice/completeplus.php

the login form above will be shown, and the user guest, password guest, can be used to login and view the file. Every "private" file has details in its header section to check whether the user is logged in and has permission to view the file in question. However, by augmenting the URL with login details, the result will be an autologin, with no login prompt – the document will be immediately displayed. The URL would then be:

<file URL>?username=guest&password=guest

So, in the example above the auto-login link would be (try this to see it in action):

https://www.pdf2html5.com/pdfupload/server/php/uploads/admin_qctcrbtsos/alice/completeplus.php?username=guest&password=guest

Note that this form of autologin will not work as an iframe link on a third-party website due to cross-domain security checks. However, such connections can be enabled with a slightly different form of link (see further, embedding, below.

In addition, for selected Webdoxx services, autologin can be enabled for specific IPAddresses or wildcard-defined blocks of IPAddresses, for example, 123.123.456.* and even 123.456.*.* - typically this is defined for users in a Custom field when the user is added – it is used mainly for access by large institutions, like Hospitals, Universities and Governmental bodies.

User Management and Access Control

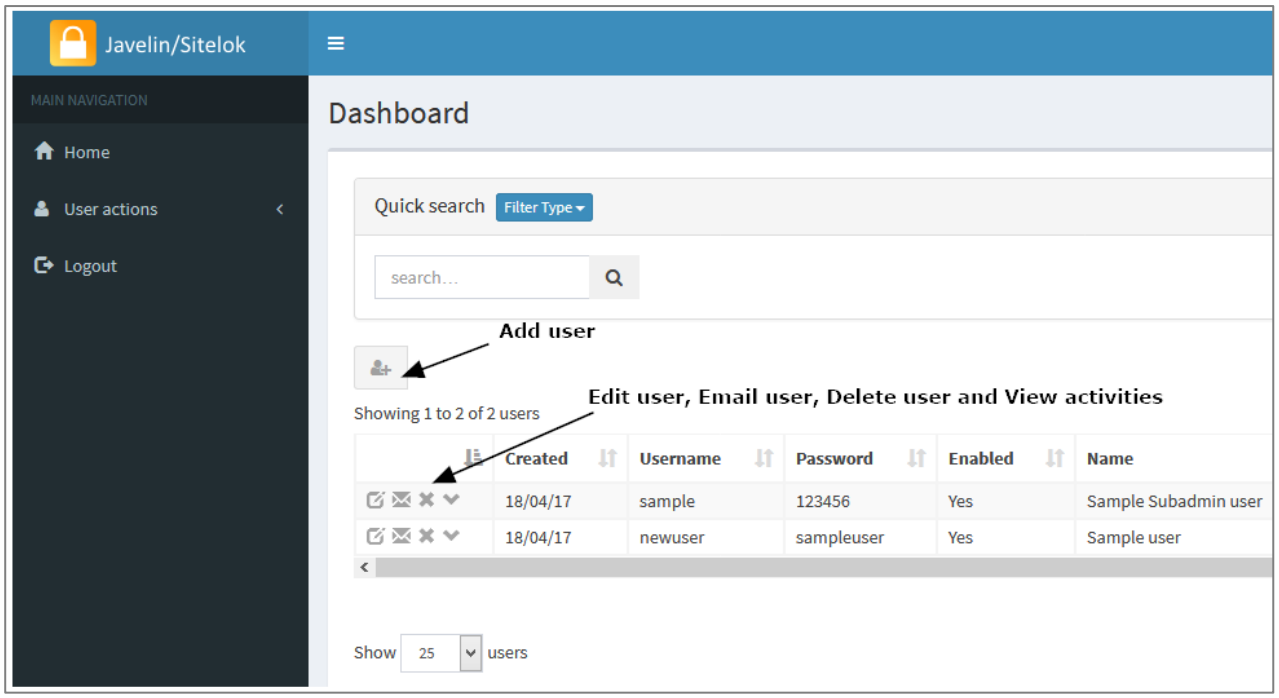
Access control and Tracking with Javelin/Sitelok

Public documents are those with no access controls provided on the Webdoxx server, so simply providing end users with the relevant link will enable them to view the document without any requirement to be registered on the service. This is useful for some publications, including: public consultation documents, free issues of magazines and newsletters, sample extracts of publications, proposals for consideration by specific interest groups (e.g., union members, public interest and lobbying groups, etc.).

In some cases, documents can be made available for private access even though they have no access controls specified on the Webdoxx servers. An example is where access to documents is provided on the publisher's own website or portal, which users have to log into for a range of services. Within that third-party website (e.g., a separate subscription or members-only service) the Webdoxx documents can be provided via an iframe connection, where the URL used is hidden and not accessible and the documents appear embedded within the publisher's own service - see further, the section on embedding, below).

Private documents are protected by access controls and these are typically provided via username/password entry. Other access control options are available, including the kind described above for "public" documents, access control with autologin via IPAddress or IPAddress blocks, and access with additional security using CAPTCHA and/or 2FA checks.

On the Webdoxx PDF2HTML5 service subscribers are allocated SUBADMIN status, together with one or more Usergroups that they are able to assign to their own documents. SUBADMIN users access these facilities via the USERS menu item, which will then show the Javelin/Sitelok simplified user management dashboard. A screenshot of the "SubAdmin" user management facility is shown below; the functions shown are all that most will need:



Access to a specific document or documents for a specific user is controlled by:

- providing the user with the specific URL for that document (directly or via a menu on a web page, or via an iframe or page re-direction that has the link defined within it) - see further, the section on embedding, below
- defining whether the document itself is set for PUBLIC access (no login required) or PRIVATE access (login required). Private access is defined and controlled by what group or groups of registered users are permitted to access that particular document. In Webdoxx, these groups of users are called USERGROUPS. The Usergroup setting for a converted private document can be specified as ALL (the pre-defined default), which allows access by any logged in user, or restricted to a specific named USERGROUP, e.g., TEST01. This Usergroup name must be specified for the document in question - this can be enabled manually (e.g., by our team) or automatically based on the publisher's registered details - selection of the Usergroup or Usergroups associated with a document on the Webdoxx PDF2HTML5 service is made via the FILES menu File Management facility for logged in Corporate and Enterprise users. If the registered user is assigned as a member of Group TEST01 then they will be permitted to view that document and any other documents assigned to Usergroup TEST01, otherwise access will not be permitted even if they are logged into the service for other document accesses
- users can also be instantly enabled/disabled, and/or have date/time restrictions placed on their Usergroup membership so that their access to documents assigned to those particular Usergroups automatically expires on a specified date

Usergroups are created by the overall System Administrator - in general this is carried out by our team or the publisher's own team, depending on the service level you subscribe to. Each of your users can then be associated with the Usergroup or Usergroups that you specify for them and as a result, will potentially have access to all documents that are members of that Usergroup.

The screenshots below show some of the Javelin/Sitelok web-based administration facilities for publishers who wish to manage their own user registrations for access control. This is a session-based security facility, with many selectable options. The screens show:

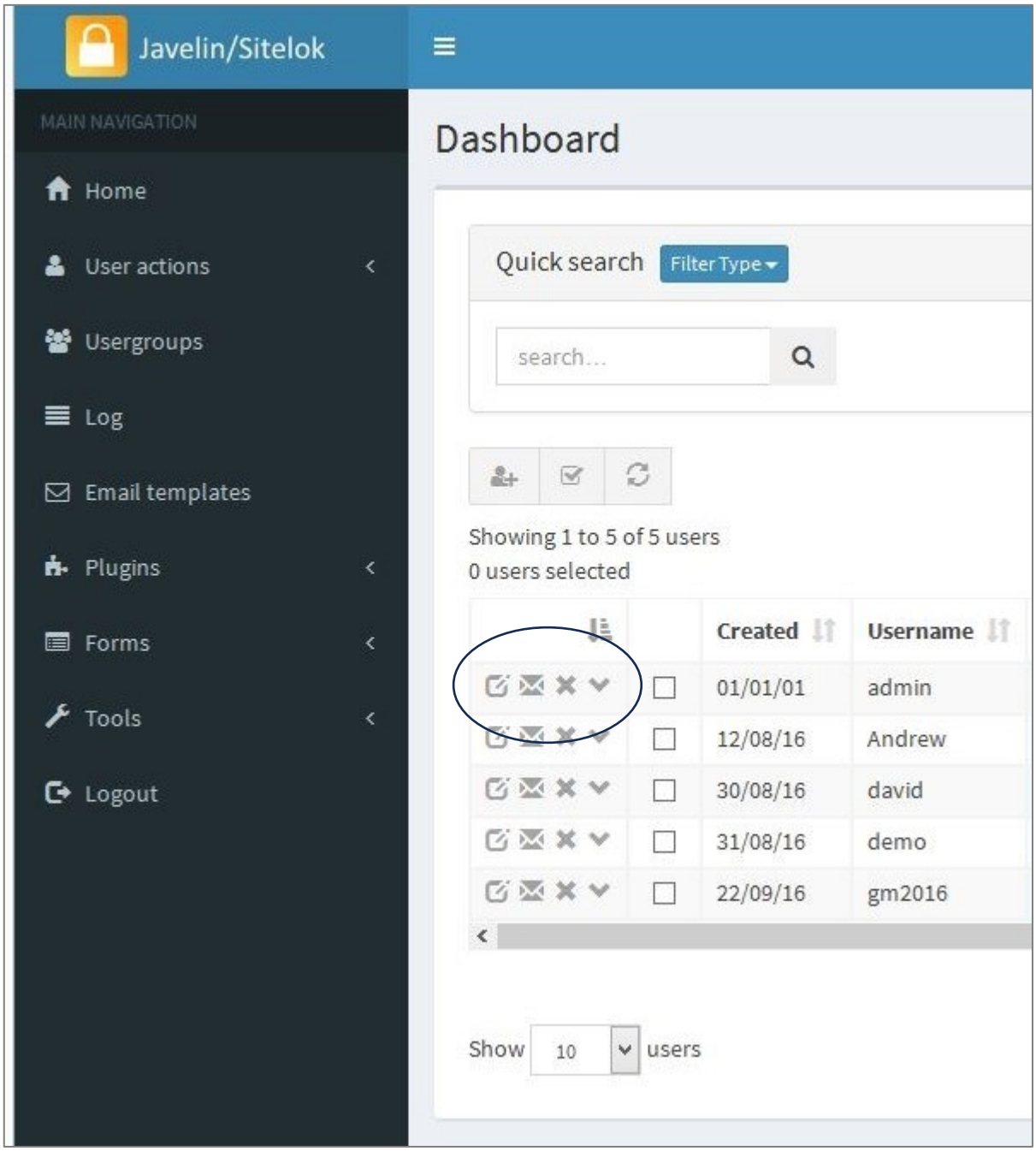
(i) the main Dashboard and function menus, with details of registered users - this is the full User Management Dashboard - the reduced version illustrated above is provided for most subscribers to the Webdoxx PDF2HTML5 service and basic managed Webdoxx service as standard (the full version is only available for dedicated service subscribers on our managed Webdoxx services)

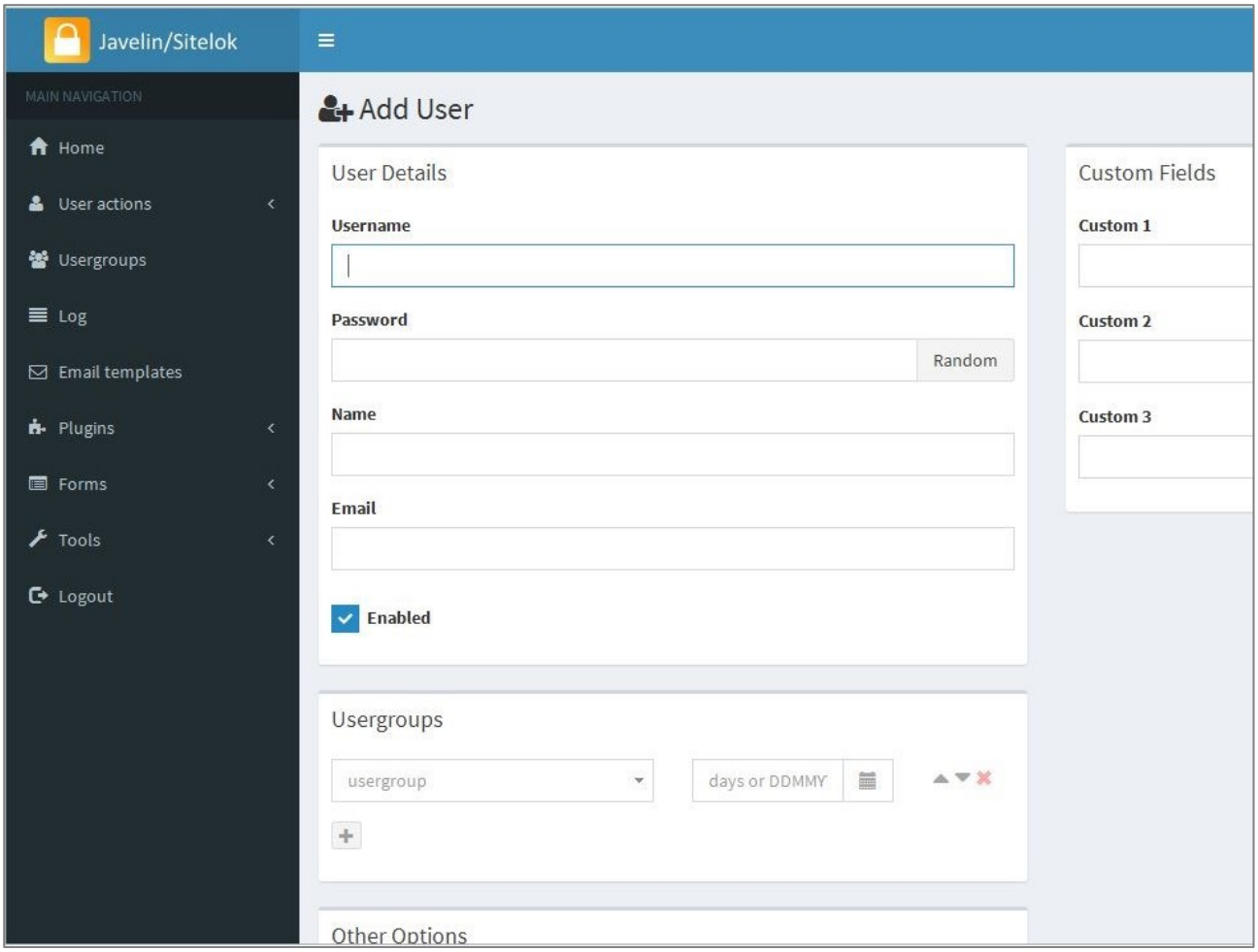
(ii) the Add User facility, where users can be manually registered one at a time, plus a screenshot of the Import Users facility. Importing of username/password lists and other user-related data is supported for fast loading of large userlists and is available as a bespoke service and for dedicated Enterprise services; and

(iii) a screen showing some of the log file details that are collected - log data can be viewed and exported for separate analysis. The User Management section of the Javelin/Sitelok manual is available on request - note that many advanced features

of the user management software are only available for bespoke/managed service projects, typically handled via our managed Webdoxx services. The facilities provided on our automated PDF2HTML5 service are a subset of those described in this extracted section of the manual. IMPORTANT: When adding a new user, you must assign them membership of at least one of the Usergroups that have been assigned to you in order for you to be able to edit and manage that user in the future. Ask us for assistance if you have any problems in setting up users and the associated Usergroup membership.

The small icons to the left of each user entry circled in the screenshot below allow the following functions: Edit user, Email user, Delete user, and view user Recent activity:





The user management system includes features for self-registration by end users ("members"), amendment of user profiles by end users, guest access and more. These can be made available within the "members" area of the service to meet bespoke requirements as part of a broader project requirement.

The default service arrangement permits multiple logins on the same username/password, with all login events tracked. Auto-login via static IPAddresses can be provided if required. Unique (non-concurrent) service access can be provided via our managed Webdoxx services, using a dedicated user management database and console.

Note: the concept of concurrent and non-concurrent logins is far from straightforward in a web-based application environment, as in almost all cases users do not log out of a service, they simply exit the browser or current tab, or just leave themselves logged in. The Javelin/Sitelok system with non-concurrent (unique) logins works in a similar manner to banking systems and if a second valid login is made, e.g., on a separate device, the first login event is automatically logged out. The Webdoxx PDF2HTML5 site allows for concurrent logins using a single username and password, i.e., does not auto-logout the previously logged in user. This has many advantages as it allows for typical end user behaviour, allows for auto-login service access from third party websites, and enables groups or classes to login using a single username/password. However, in addition, access can be restricted to a specified number of devices if so desired. This feature is enabled by setting a value in the Concurrency field for the user record in question. This could be 10, for example, which would

allow access from up to 10 devices. Each time a user logs in with that username on a particular device the Concurrency counter is reduced by 1 and a browser cookie set on that device to indicate that it can login to the service. When the Concurrency counter drops to 0 no more new devices (browsers) can be used to login with that username/password unless the user record is amended to increase the count once more. If you decide to use this feature you MUST inform your customers that your service uses cookies. Please see here for guidance from the UK authorities on compliance. Other concurrency models are also available on our managed Webdoxx platform, so arrangements can be tailored to a publisher's requirements.

Embedding

iframes

There are various ways you can provide a target document embedded on your website. Typically, this would be using an iframe, and can be with or without autologin. To start with a simple test, try providing a test document on your website, e.g., using the following example iframe snippet:

```
<iframe src="https://www.pdf2html5.com/pdfupload/server/php/uploads/sample_mfhztwpdfp/advance-coaching-pbk/complete.php" width="100%" height="800px" allow="fullscreen"></iframe>
```

An example using this snippet is provided on our Webdoxx.com website, here:

https://www.webdoxx.com/index-book.php

Typically, you would include this on a fairly empty web page, accessible only to logged-in users. In addition, you would typically "hide" the link for security reasons. There are many ways of achieving this, including using URL-shortening (e.g., bit.ly), JavaScript 'encryption' of the link, and indirection (where the URL is determined by a variable that is stored in an inaccessible file or table off the active page).

The link shown above is to a "public" document, but by adding the following to a "private" document, the connection can auto-login to the target page (not generally for use within iframes as it may result in a CORS control error – on a bespoke basis this issue can be overcome however, using the Access-Control-Allow-Origin facility – please contact us for more details):

```
?username=XXXXXX&password=YYYYYY
```

Pop-up window

If the use of the additional command line elements shown above fail to work as expected - a common issue due to cross-origin security checks (so-called CORS controls) - another approach is to use a pop-up window with autologin. Here is example JavaScript code:

```
<script> function openwindow() { window.open("<full URL of target, optionally with username and password details>","myWindow","menubar=0,status=0,scrollbars=1,resizable=1,height=1000"); } </script>

<div> Please click <A href="javascript: openwindow()">HERE</A> to access the service </div>
```

Subscription scenario

This scenario shows how you can use the service to offer a subscription service to your own customers and control their access. This is just one scenario and others may be more appropriate depending on the way in which you decide to manage your target customers and the various documents you wish to make available on a subscription payment basis:

- A company called ABC Inc with 100 branches subscribes to your publication(s) service, an annually updated Manual, on March 1st 2023. The subscription is for 12 months. You register a single new user, abcuser, with a concurrency count of 100 devices and membership of your usergroup "ABC". You specify that their membership of the ABC usergroup is set to expire on 28th Feb 2024
- you upload the 2023 manual and set the Usergroup for this publication to ABC via the FILE menu (file management facility)
- you provide ABC Inc with the link to your publication and the username and password for their organization (this could be more than one username/password of course, e.g., one for head office with 10 devices permitted, plus another for their 100 branches with perhaps 120 devices permitted - to allow for device changes in the branches). Note that the link could be provided via email or via your own website (and could be "hidden" on your site, with or without auto-login to the service, or even embedded on your site).

ABC Inc and their branch network use the service and all is fine, but then you issue the 2024 manual in January 2024. As with the 2023 manual you upload it, use the FILES menu to set the usergroup to ABC and let all your clients (including ABC Inc) know that a new version of the manual is available and provide them with the new links. ABC Inc is just one of these clients and they let their users/branches know the new link (or this is handled automatically via your own website links/iframe setup), and continue as before

ABC Inc fail to pay for the next year's subscription and think they can continue using the service, but from 1st March 2024 they will not be able to because their usergroup membership has expired.

Another subscriber, DEF Inc, say, has done exactly the same but subscribed in Jan 2023, does pay for the 2024/5 subscription so you change their group expiry date to 1st Feb 2025 and their users will be able to access the 2023 and 2024 manuals throughout the period to the end of their 2024/5 subscription

ABC Inc relent and pay up - you amend their subscription expiry date and their access springs into life again for all their users and both versions of the manual

In addition to the above, you can simply enable/disable any user at any time, for example if you detect misuse of the service or non-payment of a monthly subscription within a year' subscription window

E-commerce scenario

It is very easy to integrate our offline solution to any ecommerce platform to facilitate book sales, for example - integration of online publication sales are slightly different.

With our offline system there is no need to handle user registrations or to manage users who are sold digital products. All that is needed is a hosted block of authorization codes that are specific to each document - effectively like a list of license files or keys, but these authorization codes are simply text strings and held in a text file for each title. Every time someone places an order the next available (unused) string is picked and an email is sent to the purchaser with the download instructions link and authorization code. The code is specific to the that document and has a usage count associated with it, so typically this would allow usage of one or more devices.

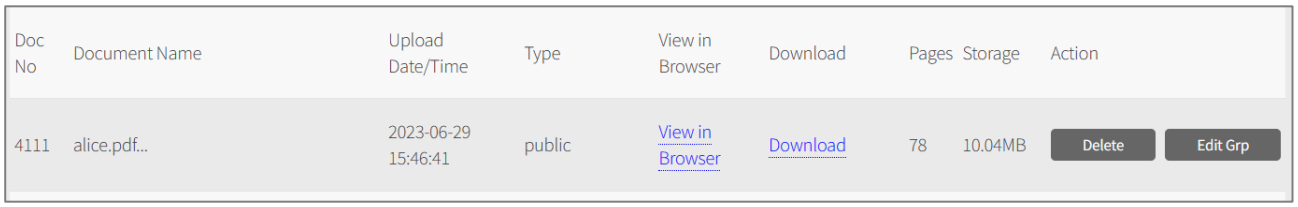
A similar arrangement can be provided for online services, but in this case the online system typically would have program-generated usernames and passwords pre-loaded into a dedicated publisher's user management database on our server, often with usergroup membership set also. The ecommerce system then picks the next available username/password combination for the publication(s) in question from the matching list held on the ecommerce site and sends that to the user. Some publishers handle this step manually, some do it via their ecommerce system, and some handle the service on their own "membership" system with users accessing our online service via their own portal or platform (linked or embedded).

Sessions and logins/logouts

The Webdoxx services use session IDs to identify logged in users. Whilst a browser is open and the user remains logged in, they can access the document or documents for which access has been enabled without repeatedly having to login for each document or after closing and opening tabs, UNLESS either they explicitly logout via a logout button, or close their web browser, or a logout is forced upon them as a result of a timeout or programmatically generated logout event. A timeout occurs if there is no activity for 7200 seconds (120 minutes) on the Webdoxx PDF2HTML5 service, but can be specified as a different value on Webdoxx dedicated user management installations. Ask us for details if you think you need to use any of these facilities.

File Management

The Webdoxx PDF2HTML5 service includes a File Management facility, accessed via the main menus when a service subscriber is logged in. The screenshot below shows a sample file listing for the logged in user "sample". Below the screenshot we explain the key features and functions provided. On managed Webdoxx services file management is managed by our team and/or via providing direct FTP access to the publisher's dedicated file storage folders on our servers.



View in browser

The file listing shows all the documents uploaded and converted by the currently logged in publisher, in date order (most recent first). Links are provided to enable the converted files to be viewed in a separate tab. The file listing page shows the number of pages in the document, whether it is public or private, and the storage utilized for each line item, together with the total storage used by all items added together (important for specific service levels - exceeding the allocated storage may result in additional charges). When files are uploaded on the Webdoxx PDF2HTML5 service they are converted to HTML5 format and stored in a folder as the converted files. Each time a file is uploaded it is assigned its own unique location and hence its own unique web link (URL). For bespoke services the re-use of existing URLs for file updates can be provided. As noted earlier, source PDFs are not retained - they are deleted immediately after the conversion process has completed.

Download

For selected subscribers the option to download a zipped fileset of the HTML5 converted files is provided (this is a "Pro" service option). This fileset can then be unzipped and then transferred to our managed Webdoxx service for display to end users. Where filesets are transferred to our managed Webdoxx services the publisher downloads and unzips the fileset, uploads them via FTP to their dedicated folder on our server, and includes an index.php file to the fileset before viewing. This (editable) index file is provided by our team and includes the user access controls that the publisher wishes to apply, the document viewer code, and other options and controls that are included at the publisher's request (e.g., branding, accessibility settings, bespoke watermarking, links to the publisher's website etc.).

Delete files

The Delete button enables all storage associated with the selected document to be deleted from the server, thus removing access to the document in question and releasing the storage space used. It is good practice to delete files no longer required, particularly if you are

approaching your agreed storage limit. Files generated by publishers who are not subscribers or are former subscribers will be deleted automatically or manually by our team.

Edit group

The Edit Group button enables you to associate one or more Usergroups with the selected document. As noted earlier, access to a specific document or documents for a specific user is controlled by:

(i) providing the user with the specific URL for that document (directly or via a menu or via an iframe that has the link defined within it or via a page re-direction); and

(ii) by defining whether the document itself is set for PUBLIC access (no login required) or PRIVATE access (Login required). For PRIVATE files, access can be further controlled by specifying what Usergroup or Usergroups of registered users are permitted to access that particular document. The Usergroups setting is specified as ALL (access allowed for any logged in user) as the default. However, in many instances it is preferable to restrict access to a specific named Usergroup, e.g., TEST01. Selection of the Usergroup or Usergroups to be associated with a document is made via the EDIT GROUP button for logged in Corporate and Enterprise users. If the registered user is assigned as a member of Usergroup TEST01 then they will be permitted to view any document that has been assigned to the Usergroup TEST01, otherwise access will not be permitted. Bespoke templates may have one or more Usergroups pre-loaded into their specification, so in this case separate Usergroup assignment for documents would not be required.

Usergroup membership

Usergroups are created by the overall System Administrator - for Webdoxx PDF2HTML5 services this is carried out by our own team. Initially a single unique Usergroup is assigned to each subscriber, but additional subscriber-specific Usergroups can be provided on request (e.g., myCourse1, myCourse2, myFinanceDept, myMarketing... etc). Each of your end users can then be associated with the Usergroup or Usergroups that you specify for them and as a result, will have access to all documents that are members of that Usergroup.

Controls that limit access to a specific user are possible, though rarely used. This can, of course, be handled via the Usergroups mechanism, with the user or users in question being the only current member of a specific Usergroup. User details can also be built-in to the template used for accessing a document. Currently this is provided as an "on request" basis, and should include the list of registered usernames that access is to be restricted to.

Example scenario: for example, a user-specific draft legal agreement document could be assigned membership of Usergroup: "**myagreements**" for a specified period and the specific link for that document and for that user sent to them (the user would have to be registered, enabled and a member of the Usergroup **myagreements**). They would access that link (which other users would not know) and login with their username and password to view the document. At some point the Agreement would be accepted (or rejected) and no doubt

a separate legal signed document or documents created with the user. The Webdoxx document in question can then be deleted from the server (unless still required) and this user (and any other "old" users) disabled, so even though they remain as potential users on the service, they would have no access - or their access could be retained and enabled, but membership of the **myagreements** Usergroup removed (manually or automatically on date expiry).

Security

Overview

There are multiple elements to the security framework in our Webdoxx service offerings:

- conversion of the source PDF to HTML5 with page separation, text separation and image or scalable vector graphic (SVG) page element rendering
- usage of different viewer formats and security measures reflecting differences in target customers (e.g., for some, copying of text might be permitted, whereas for others and by default, that would not be allowed)
- disabling text selection and copying, printing and right-click mouse operations
- optional encryption of the viewer access page using industry-standard JavaScript-based methods
- provision of session-based secure access to each document using a server-based user management framework with a range of concurrency controls
- overlaid user-specific dynamic watermarking (see further below)
- usage tracking by user and document
- https-based hosting
- optional iframe-based or pop-up form embedding
- optional pre-watermarking of source PDFs
- controls over direct access to resources
- non-retention of uploaded PDF files; and
- optional IPAddress-based access controls

These elements provide a satisfactory level of content protection for a large number of applications and documents, and can be augmented in a number of ways. For stronger security and optional secure printing, we recommend using our offline Drumlin DRM service with Javelin3 secure PDF readers.

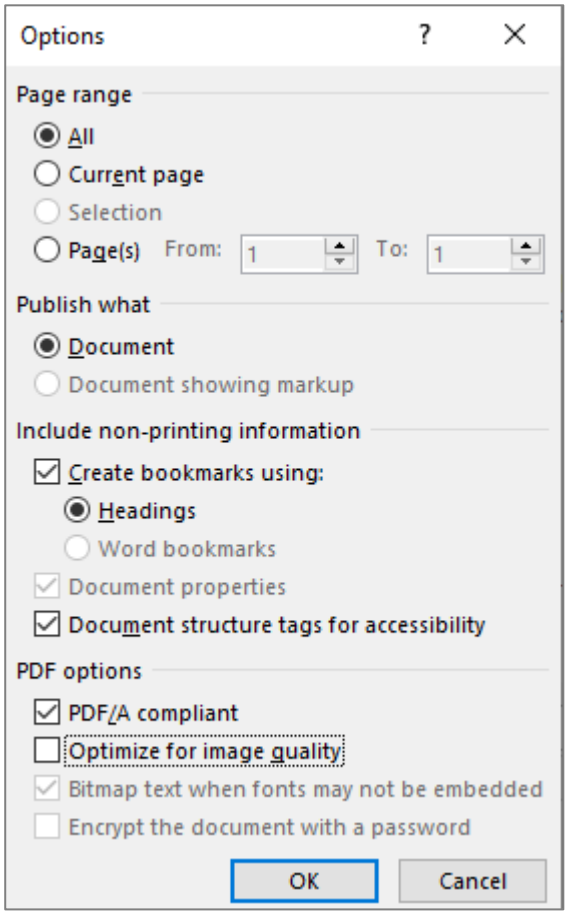
Watermarking and Bespoke templates

PDF files can have a static watermark added to selected pages or all pages before uploading to the service. A "tiled" watermark on every page adds a significant level of additional security to the source file. Dynamic watermarks are displayed on the web page, as an overlay, as standard for all Private pages. The content of the default dynamic watermark includes currently logged in user's registered name, the current date and time, and the IPAddress used to access the page. The size, font, color, opacity, content and orientation of dynamic watermarks can be amended from the standard version for bespoke display templates. For example, dynamic watermarks could include the user's company name, orientated at 45%, a large pale grey font. The filename of any such bespoke template is included in the SubAdmin user record for Webdoxx PDF2HTML5 service users, or as a modified index file for managed Webdoxx service users. For the automated service this additional template option is included in the list of links provided in the email sent to the publisher after a file has been converted, alongside the standard set.

Office and other format source documents

Word format and other text source documents

Word, Excel or PowerPoint and similar files are typically first converted to PDF using the Export to PDF facility in the MS Office suite - this produces the best/highest quality results. The converted files can then be uploaded as PDFs and displayed in HTML5. As noted earlier, please refer to the recommendations in Appendix 1 when generating new PDF files. When exporting to PDF within MS Office products like Word, we recommend using the Microsoft "Create PDF/XPS document" option, and make use of the OPTIONS button before exporting, as shown below (note that the PDF/A compliant option is ticked, although this is not essential):



PowerPoint files

On our managed Webdoxx services, files can be converted by our team on your behalf. For PowerPoint files conversion direct to HTML5 can be provided, with full support for animations, transitions and audio/video elements - for a simple interactive PowerPoint sample using secured HTML5 please click here. Contact us for more details.

ePUB files

Documents that are formatted as ePUB files can either be printed or converted to PDF and treated as normal for any PDF, or can be displayed within our ePUB viewer. For examples, please contact us.

Media files (Video and Audio)

Linked media assets, such as MP4 videos, can be made available without providing direct access to the assets from the web. This is a bespoke service arrangement and assets will contribute towards total storage allocations. Alternatively, publishers can embed video material via secure streaming provision using services from providers such as Vimeo

Printing and Downloading

Printing

Printing of the displayed document is always disabled as the default. Special arrangements can be made for PDF printing if required. Our offline Drumlin DRM-based security using our special Javelin3 PDF readers is recommended in this case.

A Print icon with Print request email, or a link to order a printed copy of a publication, can be provided as a toolbar icon. Immediate print can be enabled using our dynamically displayed PDF viewer - a sample of the dynamic display version can be viewed here (with no printing permitted!). In general, we do not recommend enabling printing for online services because the print functionality is managed by the browser software, so printing to a PDF cannot be prevented, thereby removing the whole purpose of protecting the document.

Downloading

Downloading of the source PDF is prevented by default and because no source PDF is stored on our servers this provides an additional level of security. Of course, if a dynamic display viewer is used it has a copy of the source PDF on the server, so making that available as a download is, indeed, possible, although again, not desirable in most situations.

Sample documents

An exception to the cases above applies where a sample document or non-sensitive document is made available for downloading and/or printing. In this case the document can be made available as a normal PDF for users to download, much as we provide for our sets of service documentation (see our documentation page here).

Markup and Annotation

The Webdoxx viewer has support for annotations in the source PDF such as Links, Widgets, and Popups. RichMedia, Movie, Video and Sound. Such annotations are supported (i.e., included in the file conversions) provided they use an HTML5 compatible media format (e.g., mp4). The viewer can also display nearly all annotation types including Text, FreeText, Line, Square, Circle, Polygon, PolyLine, Highlight, Underline, Squiggly, Strikeout, Stamp and Caret and Ink. Displaying of complex annotation types such as 3D are not currently supported.

Editing and resaving annotations are not currently supported. Markup and notes annotations can be enabled using third party services like Weava and Hypothesis.

Appendix 1 PDF Recommendations

If you have control of the creation of the PDF, there are some things you can do to ensure you are future-proofing the content and getting the best out of Webdoxx:

1. Avoid tools or settings that compress the PDF excessively
2. Ensure the PDF has no added Adobe-style "security" set
3. Ensure fonts are embedded
4. Enable marked/tagged/structured content
5. Create the files as PDF/A
6. Avoid Adobe-style Forms and Adobe-specific specials (e.g., embedded 3D models) as these are not supported

Avoid tools or settings that compress the PDF excessively

Tools that compress PDF files often get judged by how well they reduce the file size, and often achieve this by removing important information that can cause problems down the line. A compressed PDF can often 'look' fine, but under the hood be a different story.

Some examples of problems resulting from compressed PDF files that we have seen include:

- broken extraction of text due to the removal of character mappings
- fractional white lines appearing in images due to images getting tiled
- fractured text output due to the removal of width data in the font
- loss of image quality due to images getting overly compressed

A compressed PDF file rarely affects the file size generated by Webdoxx; therefore, we generally recommend avoiding those tools/settings if possible.

Ensure fonts are embedded

PDF files can be created to rely on fonts that are stored on the local file system instead of embedding them within the PDF file. When this happens, Webdoxx substitutes any non-embedded fonts with open-source fallbacks. To ensure the appearance remains accurate, we recommend embedding all fonts if possible.

Enable Marked/Tagged/Structured content

A standard PDF file does not contain any kind of structural information (such as paragraphs, headings, etc). Marked content is an optional feature for tagging the content in PDF files with additional structural information. Many PDF files that we see do not include it, but if you have control of the PDF creation then we do strongly recommend enabling it.

Even after a PDF has been created it is possible to improve it significantly using tools like Adobe Acrobat Pro. With these tools you can do things like:

- Ensure no Adobe-style security has been set (e.g., password to open, settings that define the print and view permissions)
- Automatically generate a Bookmarks (Outline) tree (recommended for any large document)
- Amend the pagination for a better end-user experience (often helpful, for example removing blank pages, or adding a blank page to ensure two-page spreads appear correctly/as expected, where these are the default form of presentation)
- Add a Title to the Document Properties
- Convert text that are implicit links into explicitly defined links (e.g., to websites or email addresses)
- Ensure any Contents page is linked to the sections it refers to
- Optimize the structure of the document for accessibility
- Remove any unwanted print markup elements
- Crop the pages to a single standard size – for example, to ensure that a mix of 1- and 2-page layouts are not used, and ideally ensure that all pages are in the same orientation (portrait or landscape, rather than a mix)
- Add media-related links, e.g., to audio or video clips
- Clean up any errors in the presentation of the file
- Save the file to comply with standards, notably PDF1.6 or earlier, or PDF/A (see below)

Create the files as PDF/A

PDF is a very powerful file format, and with great power comes great responsibility. Not all PDF creation tools are equal, and some do a better job of it than others. As with HTML parsers, PDF parsers are expected to handle documents that do not fully comply with the specification. We are often making tweaks to our parser to handle documents with questionable interpretations of the PDF specification.

Enter PDF/A: PDF/A is a more modern, stricter version of the specification that includes provisions to ensure the document preserves information relating to content extraction and document accessibility. This goes beyond the intentions of the original PDF specification which was primarily as a print format.

Accessibility

There are various ways of implementing audio versions of books without resorting to an actor reading out all the text. A number of web browsers support text to speech as standard (e.g., MS Edge, Safari on the Mac etc.) but there are many plugins that aim to provide this kind of functionality for different browsers. Broadly speaking they are of variable quality and in general require the ability to select the text you want to be read, otherwise what they try and read out can be pretty random and broken up into unintelligible chunks.

We have experience in ensuring documents are pre-processed to optimize their accessibility, including ensuring the structure, tagging, linearity etc. of the content is as accessible as possible. If text can be selected then it can be copied of course, so we do not normally allow this although it can be permitted if required/needed for screen readers like NVDA. The other two options (for acknowledged visually impaired students) are to provide an audio version

of the book or sections of the content (online or downloadable), or to provide a designated copy of the source PDF for use in special devices/software produced for the visually impaired. These options can provide a far better experience for such users.

Appendix 2 Language Variants

PDF documents that specify a Right-to-Left reading order (such as Arabic and Hebrew and vertical writing systems such as Chinese, Japanese, Korean) are also supported in all Webdoxx Viewer modes. The layout direction in magazine layouts is updated and the navigation buttons are swapped to match the document direction.

A wide range of languages are currently supported for toolbar elements, drop-down lists and messages, and we are looking to add more languages in the future. If you would like to help us by contributing a language translation, please let us know and we would be happy to provide you with the translation list.