Biotext

Web content format – HTML, Word or PDF?

This guide will help you to understand the best format for your content.

Format considerations

In general, HTML is the most accessible and findable format for online content. Screen readers can read the content aloud and AI models can detect and parse the content to use it in AI answers to user questions.

Many websites will also publish Word or PDF documents. These can be faster to publish than HTML, if the documents already exist. They also allow users to download and print content.

Screen readers and AI models can only work with Word and PDF files if they are fully accessible. Word files should use styles for all content. PDFs should be tagged to define headings and reading order. All images should have alt text.

Content in PDFs and Word files is generally less visible to AI models. Some AI systems will only look at the HTML content on the page. Some will look at both HTML and linked documents, but HTML content will be prioritised over separate files.

User research will help you to understand which format is best. This will depend on the type of information your users are accessing and how they want to use it. The following table outlines some of the advantages and disadvantages of each format.

Format	Advantages	Disadvantages	Good for
HTML	• Content is most findable for search engines and AI models • Easy to make responsive for viewing on desktop or mobile • Easy to update • Accessible for people with disability and those using assistive technologies • Loads faster than other formats, and can be viewed without downloading • Supports interactive content • Can use Google Analytics (or similar) to understand users' behaviour	• May not look good if printed • May be harder for users to know when the content was published or updated • Cannot be accessed without an internet connection • New content must be planned for and fitted into the existing information architecture • Appearance (text flow, image sizes and so on) will change with users' web browser choices and configurations	• Most of your web content • Interactive content • Content that needs to be updated often

Format	Advantage	Disadvantage	Good for
Word (.docx)	• Easy to produce and update • Easy to include publication dates • Users can download and/or save to their own device if they want to refer to the content often • Once downloaded, does not need an internet connection to view • Easy for users to print	• Not as easily findable by search engines or AI models • Slower to access; needs to be downloaded to view • Harder for users to navigate back to other web content and may affect the user experience • Will not integrate with other technologies or add-ons • Requires specialist knowledge and planning to ensure it is accessible • Not interactive • Users can edit it (unless it is locked) • Appearance may depend on version of Word, fonts installed and so on • Is not responsive and may look different when viewed on a mobile device than on a desktop • Google Analytics can track if it is downloaded, but you won't gain any other insights about users' behaviour • Can result in developers uploading a file because it is more convenient, rather than incorporating content into webpages • Old, outdated Word documents and PDFs must be archived, otherwise they can linger on a website and confuse users	• Content that users will need to print • Forms
PDF	Same as for Word, plus: • Looks good if printed, especially if professionally formatted or designed • Appearance is consistent across devices	Same as for Word, plus: • More difficult to update if content needs to change, especially if professionally designed using design software	• Long reports or documents that will not change often • Content that users will need to print • Forms