INTERNATIONAL COUNCIL FOR HARMONISATION OF
TECHNICAL REQUIREMENTS FOR REGISTRATION OF
PHARMACEUTICALS FOR HUMAN USE

ICH M2

REDACTION: POINTS TO CONSIDER

1 PURPOSE OF THIS DOCUMENT

This document provides points to consider on the topic of redaction of information from documents submitted to regulatory authorities and which will subsequently be made available to the general public.

The intent of this document is not to try and harmonise the redaction requirements that have been published by various regulatory agencies but to provide points to consider about the ways in which redaction can be achieved. The requirements of each regulatory agency should be determined before proceeding with a redaction of a document for that agency.

2 WHAT IS REDACTION?

The term “redaction” in this “points to consider” document is used with the following definition:

Redaction: To delete private or sensitive information from a document in preparation for publication.

Recently a number of regulatory authorities have produced guidance on the need to publish redacted versions of documents that they have received on their public websites. The intent of the redacted version is to remove from view information that may, for example, be of a personal nature or commercially sensitive.

In traditional paper documents, the deletion of the information is usually represented by a solid black box covering the area where the deleted information was originally presented on the page. Electronic redaction aims to delete the information from the file and this is usually represented on the screen or in the physical version of the document with a solid black box where the information was in the original file. However, in some circumstances it may be acceptable to simply delete the information and give no representation of the information in the final redacted file.

The redaction of information in electronic files also means that consideration must be given to the metadata associated with the electronic file and whether this contains private or sensitive information that should also be deleted.

This “points to consider” document will identify a number of areas to consider associated with this topic. The exact details of the requirements or acceptable ways of redacting files and documents will be subject to local guidance and regulations.

3 POINTS TO CONSIDER

3.1 General Points to Consider

3.1.1 Permanent Deletion of the Information

Redaction of information must be done in a way that permanently deletes the identified information such that the consumer of the document is unable to read or reconstruct the redacted information.

In paper documents, it is not sufficient to merely cover the text, it is recommended that the document is then copied for further distribution and consumption.

For electronic files, it is not usually enough to obscure the information in the page rendering of the information as search/copy tools can usually still identify the information. Instead, the information must be completely deleted from the file.

3.1.2 Representation of the Deleted Information

As has been noted, in paper documents, or the representation of paper pages from electronic sources, it is usual to show where the deleted information was in the original document by placing a solid black box of the same size and area as the original text. This arises from the historical method of redacting information which was to cover the text with black tape or using a thick tipped black marker pen.

In an electronic file it may be possible to ensure that the display of the information on a page representation also includes a black box. Alternatively, when deleting the information it may be possible to replace the characters with dummy characters (e.g. a string of XXXXX) to show where the redacted information was in the original file. The replacement of the original information with a black box or replacement characters allows for easier review of the redacted file alongside the original. It should also be noted that the representation of the full page may be important if there was key information in the headers or footers, or in attached annotations and notes, that were not subject to the redaction process.

It should be noted that representing the deleted information with black boxes or alternative characters may give the user of the redacted document some idea about what the original information was, especially if the redaction is only being applied to a few words. For example, if replacement characters are added on a 1:1 basis, then it may be possible to count the number of replacement characters to see how many characters were in the original text. If this is considered a risk then this should be taken into consideration when processing the document for redaction.

In some instances when redacting an original electronic file it may be acceptable to just delete the information and not have any substitution to show where the original information was located. In these instances, it should be noted that the information on the page will reflow, particularly where large amounts of information are redacted. This will lead to a redacted file with a different page representation (e.g. different number of pages, different layout of information, etc.) to the original and it may also lead to apparently nonsensical sentences if individual words or phrases have been removed and no indication that this has taken place is noted.

In general, it is recommended that a file that has been processed for redaction is marked up in some way so that this is conveyed to the consumer of the document. They can then expect that issues in understanding the information may be attributable to the redaction process.

The method of redacting information may be subject to local rules and regulations and this should be taken into consideration when selecting the means of deleting information.

3.1.3 Manual versus Automatic Redaction

Creation of a redacted version of a file is often done manually. A copy of the original file is made and this is then reviewed by someone who will identify the information to be redacted (according to the relevant guidance) and take actions to delete the information and, in general, provide the visual cue that they have done so (e.g., by creating a solid black box where the original information was located).

It is possible to automate redaction. It could be that the information to be deleted appears in a set position on the page, that key terms always appear in the information to be deleted or that some kind of NLP (Natural Language Processing) processing can be applied to identify information for redaction. Having automatically identified the information for redaction, the tool can then delete the information and, in general, provide the visual cue on the page representation. The user should consider the accuracy of the automation and the level of checking they may wish to carry out. Care should be taken to ensure that all the desired information has been redacted, but that none of the other information has been redacted accidentally.

With some electronic file formats it may be possible to attach a "tag" to identify information that should be redacted. The automation requires that a suitable tool can process the file to identify these tags and then take suitable actions to display the redacted information file in an appropriate way. The tagging of information in documents supports the true automation of redaction, but so far there is little experience how to handle tagged information in the redaction process.

3.1.4 Number of Files

The redaction of information potentially leads to the creation of new files, i.e. the original file and a redacted copy. In most cases this is unavoidable as the current tools and technologies available mean that redaction is only possible by making a copy of the original file and then redacting information in that copy. In these cases it is recommended that the relationship between the original file and the redacted copy is maintained in the way that the information is displayed (i.e. that the area that the redacted information displays is the same as the area of the information in the original document, rather than reducing the length of the redacted copy of the information). However, it is also important that this is done in such a way that the consumer cannot access the original file.

As noted above, there are ways in which the information that will eventually be redacted in a file can be marked up or “tagged” at the time the content is created. This will allow for the creation and management of a file with metadata that can control the way the content is displayed to consumers. This approach relies on both the author and the consumer having access to the necessary tools to allow the redaction processing to take place. Furthermore, the processing of the file to redact information should create a copy file that is then distributed to other consumers to ensure that the information has been deleted and is not still within the redacted file.

Note that changing the original file (if allowed) will invalidate any checksums associated with the original file. However, it does not invalidate the purpose of the checksum to validate the transmission of the original file from sponsor to regulatory agency.

3.1.5 Redaction of Metadata

Increasingly, the information that a user consumes from an electronic file will include metadata that is stored with the content of the file that is displayed. If this metadata meets the requirements for redaction in any given situation (i.e. it contains commercially sensitive or private information) then this metadata should also be processed when redaction takes place.

The same basic considerations apply to redacting metadata as apply to redacting information. Consideration should be given to replacing the metadata with dummy characters or text to ensure the consumer of the metadata knows that information has been deleted.

It should be noted that the filename is metadata and may contain information that needs to be redacted.

In addition it may be important to consider the requirements for the metadata within the file, particularly whether certain metadata values are required to produce valid files. In this case it will be necessary to know whether the consumers of the information can be informed if the file is technically invalid as a result of the redaction process but still usable in the user setting. If the technical validity of the file remains a key requirement then the redaction process must create not only redacted information and metadata but it must also represent the redacted metadata in a way that retains the validation status of the file. For example, if the metadata must have a value, then simple deletion of the text would produce an invalid file. Therefore, the redaction must replace the deleted metadata value with some characters that retain the validation status of the file.

3.2 Points to Consider for Specific File Types

The following file formats addressed in this document are all the subject of ICH ESTRI technical recommendations.

3.2.1 Portable Document Format (.pdf)

Care must be taken to ensure that information in the PDF file is deleted from the file, not just hidden. Therefore, it is not recommended that the user use a tool like the “Highlight text” tool and apply a black highlight to the information. This is because PDF copy and search capabilities will usually continue to allow the consumer to identify and select text that has been covered by an obscuring shape or changed font color to match the background. If there is any doubt, select the “Text Select” tool and click and drag over the area of the redacted text to see if the original characters can be selected.

Some PDF tools have a redaction functionality included to allow the user to manage the deletion of information.

The Document Properties/Document Information Dictionary of a PDF file are metadata and should also be considered for redaction processing.

Note that the PDF/A format usually prevents the user from making changes to a file and saving them, so PDF/A is not generally suitable as a starting point for the preparation of a redacted copy of a PDF file. If the source PDF file for redaction is in PDF/A format it must first be converted to another PDF format before any redaction processing takes place. The following two links may help provide information about how to change a PDF/A document so that it can be edited (links retrieved on 15-May-2017).

- https://ord.uscourts.gov/index.php/about-cmecf-and-pacer/cm-ecf-help-all-help/201-pdf-help/559-editing-pdfa-documents
- http://blogs.adobe.com/acrolaw/2011/05/how-to-remove-pdfa-information-from-a-file/

3.2.2 MS Word (.doc and .docx)

In general, Microsoft (MS) Word files from before Office 2007 and before do not offer any particular advantages for processing for redaction of information. If the provision of source MS Word files is acceptable under regional or local regulatory requirements then the general considerations for redaction that have been noted above will apply.

MS Office 2007 offered the capability to save MS Word files to the Office Open XML (ISO/IEC 29500) specification, most easily identified by the use of the .docx file extension. The OpenOffice specification allows the author to tag information at the time of authoring and for the output file to manage these tags for processing like an XML file. In theory, this would allow the author to tag information for automated redaction, as noted in Section 3.1.3. This method of working relies on both the author of the document and the consumer having the technology and processes in place to allow for the automated redaction of information. At the time of writing of this “points to consider” document this capability has not been developed by any region or local country for use in the pharmaceuticals industry.

It is noted that all of the MS Office tools from Office 2007 onwards contain the option to save the file to meet the Office Open XML (ISO/IEC 29500) specification, so the same considerations could also apply to other MS Office files (e.g. MS Powerpoint files with a .pptx extension, MS Excel files with a .xlsx extension, etc.).

3.2.3 Extensible Markup Language (.xml)

The use of XML allows the author of the XML content to insert tags to identify text for future redaction at the time the content is created. In theory, this would allow the author to tag information for automated redaction, as noted in Section 3.1.3. This method of working relies on both the author of the document and the consumer having the technology and processes in place to allow for the automated redaction of information. At the time of writing of this “points to consider” document this capability has not been developed by any region or country for use in the pharmaceuticals industry.

4 FURTHER READING

http://www.nationalarchives.gov.uk/documents/information-management/redaction_toolkit.pdf

http://www.ema.europa.eu/docs/en_GB/document_library/Regulatory_and_procedural_guideline/2017/04/WC500225880.pdf