

Allegato 2

FORMATI

BOZZA

Indice

1	INTRODUZIONE	3
1.1	FINALITÀ.....	3
1.2	I FORMATI	3
1.2.1	Le Specifiche	3
1.2.2	Identificazione	3
1.2.3	Le tipologie di formato.....	4
1.2.4	Formati Immagini	4
2	CRITERI DI SCELTA DEI FORMATI.....	5
2.1	Caratteristiche dei formati	5
2.1.1	Apertura.....	5
2.1.2	Sicurezza.....	5
2.1.3	Portabilità	6
2.1.4	Supporto allo sviluppo	6
2.1.5	Diffusione.....	6
2.2	Scelta dei formati	6
3	I FORMATI INDICATI PER LA CONSERVAZIONE.....	7
3.1	Testi/Documenti.....	7
3.1.1	PDF/A.....	7
3.1.2	DOCX.....	8
3.1.3	ODT.....	9
3.2	Immagini Raster	9
3.2.1	TIFF.....	10
3.2.2	JPEG / JPEG 2000	10
3.3	Immagini Vettoriali.....	11
3.3.1	SVG.....	11
3.3.2	Shapefile	12
3.4	Audio e Video	12
3.4.1	MP3	13
3.4.2	MPEG-4	13
3.5	Altri Formati	14
4	TABELLA DEI FORMATI PER TIPOLOGIE DI USO.....	14

1 Introduzione

1.1 Finalità

Il presente documento fornisce indicazioni iniziali sui formati dei documenti informatici che per le loro caratteristiche sono, al momento attuale, da ritenersi coerenti con le regole tecniche del documento informatico, del sistema di conservazione e del protocollo informatico.

I formati descritti che garantiscono maggiormente i principi dell'interoperabilità tra i sistemi di conservazione.

Il presente documento, per la natura stessa dell'argomento trattato, viene periodicamente aggiornato sulla base dell'evoluzione tecnologica e dell'obsolescenza dei formati e pubblicato online sul sito di DigitPA.

1.2 I formati

La leggibilità di un documento informatico dipende dalla possibilità e dalla capacità di interpretare ed elaborare correttamente i dati binari che costituiscono il documento, secondo le regole stabilite dal formato con cui esso è stato rappresentato.

Il formato di un file è la convenzione che viene usata per interpretare, leggere e modificare il file.

1.2.1 Le Specifiche

I formati sono definiti attraverso specifiche tecniche. Queste sono coperte da brevetti industriali e sono completamente o parzialmente rese disponibili, secondo quanto descritto nelle licenze d'uso. In questa categoria rientrano sia le licenze d'uso onerose, sia quelle gratuite tipo GPL. Tuttavia, almeno per ora, si ritiene la distinzione irrilevante ai fini della scelta dei formati consoni alla conservazione. È cura di ogni soggetto valutare attraverso un'attenta analisi costi-benefici la scelta più adatta alle proprie esigenze.

1.2.2 Identificazione

L'associazione del documento informatico al suo formato può avvenire, attraverso le seguenti modalità:

1. l'estensione: una serie di lettere, unita al nome del file attraverso un punto, ad esempio [nome del file].docx identifica un formato testo di proprietà della Microsoft;
2. il *magic number*: i primi byte presenti nella sequenza binaria del file, ad esempio 0xffd8 identifica i file immagine di tipo .jpeg
3. I metadati espliciti: l'indicazione "application/msword" inserita nei tipi MIME che indica un file testo realizzato con l'applicazione Word della Microsoft

1.2.3 Le tipologie di formato

L'evolversi delle tecnologie e la crescente disponibilità e complessità dell'informazione digitale ha indotto la necessità di gestire sempre maggiori forme di informazione digitale (testo, immagini, filmati, ecc.) e di disporre di funzionalità più specializzate per renderne più facile la creazione, la modifica e la manipolazione.

Questo fenomeno porta all'aumento del numero dei formati disponibili e dei corrispondenti programmi necessari a gestirli nonché delle piattaforme su cui questi operano.

In particolare, volendo fare una prima sommaria, e non esaustiva, catalogazione dei più diffusi formati, secondo il loro specifico utilizzo possiamo elencare:

- Testi/documenti (DOC, HTML, PDF,...)
- Calcolo (XLS, ...)
- Immagini (GIF, JPG, BMP, TIF, EPS, SVG, ...)
- Suoni (MP3, WAV, ...)
- Video (MPG, MPEG, AVI, WMV,...)
- Eseguibili (EXE, ...)
- Archiviazione e Compressione (ZIP, RAR, ...)

1.2.4 Formati Immagini

Per la rappresentazione delle immagini sono disponibili diversi formati, che possono essere distinti secondo la grafica utilizzata: raster o vettoriale.

1.2.4.1 Raster

Nel caso della grafica raster, l'immagine digitale è formata da un insieme di piccole aree uguali (pixel), ordinate secondo linee e colonne.

I formati più diffusi sono il .tiff (usato dai fax), il .jpeg, il .bmp. Anche il .pdf può considerarsi un elemento di questo insieme: ne fa testimonianza l'uso elettivo da parte degli scanner oggi in commercio.

1.2.4.2 Vettoriale

La grafica vettoriale è una tecnica utilizzata per descrivere un'immagine mediante un insieme di primitive geometriche che definiscono punti, linee, curve e poligoni ai quali possono essere attribuiti colori e anche sfumature.

I documenti realizzati attraverso la grafica vettoriale sono quelli utilizzati nella stesura degli elaborati tecnici, ad esempio progetti di edifici.

Attualmente i formati maggiormente in uso sono:

- DWG, un formato per i file di tipo CAD, sviluppato da Autodesk come database di definizione del disegno per il software AutoCAD, di cui non sono state rilasciate le specifiche;

- DXF, un formato simile al DWG, anche questo sviluppato da Autodesk, ma con la differenza che quest'ultima ne ha rilasciato le specifiche tecniche
- Shapefile sviluppato dalla ESRI: un formato vettoriale per sistemi informativi geografici.

Il formato ha lo scopo di rendere interoperabili i sistemi ESRI con gli altri GIS. Di fatto è diventato uno standard per il dato vettoriale spaziale, e viene usato da una grande varietà di sistemi GIS.

Un discorso a parte merita SVG. Si tratta di un'estensione dell'XML, in grado di visualizzare oggetti di grafica vettoriale. Alcuni paesi, tra cui la Francia, l'hanno inserito nei formati adatti alla conservazione documentale.

2 Criteri di scelta dei formati

Ai fini della conservazione, è necessario scegliere formati che possano garantire la leggibilità e la reperibilità del documento informatico nel tempo.

2.1 Caratteristiche dei formati

Le caratteristiche di cui bisogna tener conto nella scelta sono:

1. apertura
2. sicurezza
3. portabilità
4. supporto allo sviluppo
5. diffusione

2.1.1 Apertura

Un formato si dice "aperto" quando è conforme a specifiche pubbliche, cioè disponibili a chiunque abbia interesse ad utilizzare quel formato. La disponibilità delle specifiche del formato rende sempre possibile la decodifica dei documenti rappresentati in conformità con dette specifiche, anche in assenza di prodotti che effettuino tale operazione automaticamente.

Questa condizione si verifica sia quando il formato è documentato e pubblicato da un produttore o da un consorzio al fine di promuoverne l'adozione, sia quando il documento è conforme a formati definiti da organismi di standardizzazione riconosciuti. In quest'ultimo caso tuttavia si confida che quest'ultimi garantiscono l'adeguatezza e la completezza delle specifiche stesse.

Nelle indicazioni di questo documento si è inteso privilegiare i formati già approvati dagli Organismi di standardizzazione internazionali quali ISO e ETSI.

2.1.2 Sicurezza

La sicurezza di un formato dipende da due elementi il grado di modificabilità del contenuto del file e la capacità di essere immune dall'inserimento di codice maligno. Nel

caso dei documenti informatici dedicati alla conservazione è previsto che questi non contengano elementi che possano modificarne il contenuto.

2.1.3 Portabilità

Per portabilità si intende la facilità con cui i formati possano essere usati su piattaforme diverse, sia dal punto di vista dell'hardware che del software, inteso come sistema operativo.

2.1.4 Supporto allo sviluppo

Attualmente esistono tre modelli di supporto allo sviluppo di prodotti informatici e quindi anche al mantenimento nel tempo dei formati, quello che fa capo alle società direttamente coinvolte nella produzione, quello in cui le comunità di sviluppatori si uniscono intorno ad un progetto open source e quello di tipo misto in cui la comunità è supportata da una società ICT.

2.1.5 Diffusione

La diffusione di un formato (cioè il suo massivo utilizzo per la produzione e la memorizzazione dei documenti informatici), è un elemento che influisce sulla probabilità che esso venga supportato nel tempo. Infatti, la presenza nel mondo di un numero significativo di documenti memorizzati in un determinato formato, costituisce un incentivo a sviluppare e mantenere software per la gestione di tali documenti. Inoltre, quando un formato è particolarmente diffuso, quasi sempre sono disponibili più prodotti commerciali idonei alla sua gestione e visualizzazione, circostanza che riduce la dipendenza dal fornitore che ne detiene le specifiche incrementando le garanzie che il formato possa essere utilizzato e visualizzato senza problemi anche dopo diverso tempo.

2.2 Scelta dei formati

La scelta dei formati idonei alla conservazione è affidata al singolo soggetto.

È evidente, per quanto fin qui considerato, che si ritiene opportuno privilegiare i formati che sono standard internazionali evitando, per quanto possibile, la scelta di documenti in formato proprietario.

Ogni soggetto potrà comunque valutare la possibilità di conservare i documenti in formati aperti che, pur se non riconosciuti da organismi di standardizzazione internazionali, tuttavia, secondo i criteri esposti offrono elevata garanzia di leggibilità nel tempo.

Formati diversi, non indicati nel presente documento, possono essere scelti nei casi in cui la natura del documento informatico lo richieda per un utilizzo specifico nel suo contesto tipico dandone opportuna evidenza nella documentazione dei sistemi di tenuta e/o conservazione dei documenti informatici.

3 I formati indicati per la conservazione

I formati di seguito indicati, fanno parte di un primo elenco di formati che, con le dovute attenzioni (tra cui il controllo preventivo dell'esistenza all'interno del documento di codice indesiderato e potenzialmente pericoloso) possono essere usati per la conservazione.

Come già indicato nelle premesse questo elenco sarà periodicamente aggiornato.

3.1 Testi/Documenti

3.1.1 PDF/A

Il PDF (Portable Document Format) è un formato creato da Adobe nel 1993. E' stato concepito per rappresentare documenti complessi in modo indipendente dalle caratteristiche dell'ambiente di elaborazione del documento. Nell'attuale versione gestisce varie tipologie di informazioni quali: testo formattato, immagini, grafica vettoriale 2D e 3D, filmati.

Il formato è stato standardizzato in una serie di sotto-formati, tra cui il PDF/A per la conservazione a lungo termine.

Sviluppato da	Adobe Systems http://www.adobe.com/
Anno di rilascio prima versione	1993
Estensione	.pdf
Tipo MIME	application/pdf
Formato aperto	Sì
Specifiche tecniche	pubbliche dalla versione 1.3
Standard del PDF/A	ISO 19005-1:2005
Ultima versione	1.7
Possibile presenza codice maligno	No
Collegamento utile	http://www.pdfa.org/doku.php

Il formato PDF/A ha l'obiettivo di garantire la leggibilità nel lungo periodo evitando quelle opzioni che possono creare problemi di compatibilità di formato o di dipendenza da informazioni esterne al documento. Le caratteristiche sono:

- Contenuti audio e video sono vietati
- Javascript ed invocazioni di file eseguibili sono vietate

BOZZA

- La crittografia è soppressa
- L'utilizzo di meta-dati standard è obbligatorio
- Auto-contenuto, cioè che abbia al suo interno tutte le informazioni necessarie alla sua rappresentazione.

Un documento PDF/A può essere firmato digitalmente in modalità nativa.

Sono disponibili prodotti per la verifica della conformità di un documento PDF al formato PDF/A.

3.1.2 DOCX

Sviluppato da	Microsoft http://www.microsoft.com http://www.microsoft.it
Anno di rilascio prima versione	2007
Estensioni	.docx
Tipo MIME	
Formato aperto	Sì
Derivato da	XML
Specifiche tecniche	pubblicate da Microsoft dal 2007
Standard	ISO/IEC DIS 29500
Ultima versione	1.1
Possibile presenza codice maligno	Sì
Collegamenti utili	http://msdn.microsoft.com/en-us/library/aa338205.aspx www.iso.org

Comunemente abbreviato in OOXML, è un formato di file, sviluppato da Microsoft, basato sul linguaggio XML per la creazione di documenti di testo, fogli di calcolo, presentazioni, grafici e database.

Open XML è adottato dalla versione 2007 della suite Office di Microsoft.

Oltre all'estensione .docx, usata per i documenti di testo, esistono la .xlsx per i fogli di calcolo e la .pptx per le presentazioni.

Per garantire la caratteristica di immodificabilità del documento, è necessario controllare l'eventuale presenza di codice indesiderato (script, chiamate ad eseguibili, macro, ecc.), attraverso gli specifici tool.

3.1.3 ODT

Sviluppato da	OASIS http://www.oasis-open.org/ Oracle America (già Sun Microsystems) http://www.oracle.com/it/index.html
Anno di rilascio prima versione	2005
Estensioni	.odt
Tipo MIME	application/vnd.oasis.opendocument.text
Formato aperto	Sì
Derivato da	XML
Specifiche tecniche	pubblicate da OASIS dal 2005
Standard	ISO/IEC 26300:2006 UNI CEI ISO/IEC 26300
Ultima versione	1.0
Possibile presenza codice maligno	Sì
Collegamenti utili	http://books.evc-cit.info/ http://www.oasis-open.org www.iso.org

ODF (Open Document Format, spesso referenziato con il termine OpenDocument) è uno standard aperto, basato sul linguaggio XML, sviluppato dal consorzio OASIS per la memorizzazione di documenti corrispondenti a testo, fogli elettronici, grafici e presentazioni.

Secondo questo formato, un documento è descritto da più strutture XML, relative a contenuto, stili, metadati ed informazioni per l'applicazione.

Oltre all'estensione .odt, indicata per i documenti di testo, ve ne sono altre (.ods - .odp - .odg - .odb) indicate per la realizzazione rispettivamente di fogli di calcolo, presentazioni, grafica e database.

E' necessario controllare, prima della conservazione, l'eventuale presenza di codice indesiderato (script, chiamate ad eseguibili, macro, ecc.).

3.2 Immagini Raster

Nel caso della grafica raster, l'immagine digitale è formata da un insieme di piccole aree uguali (pixel), ordinate secondo linee e colonne.

3.2.1 TIFF

Sviluppato da	Aldus Corporation in seguito acquistata da Adobe
Anno di rilascio prima versione	1986
Estensioni	.tiff, .tif
Tipo MIME	image/tiff
Formato aperto	No
Specifiche tecniche	pubblicate da Adobe
Ultime versioni	TIFF 6.0 del 1992 TIFF Supplement 2 del 2002
Diffusione	Alta
Collegamenti utili	http://partners.adobe.com/public/developer/tiff/index.html

Di questo formato vi sono parecchie versioni, alcune delle quali proprietarie (che ai fini della conservazione nel lungo periodo sarebbe bene evitare). In genere le specifiche sono pubbliche e non soggette ad alcuna forma di limitazione.

Questo è un formato utilizzato per la conversione in digitale di documenti cartacei. Il suo impiego va valutato attentamente in funzione del tipo di documento da conservare, in quanto genera documenti informatici di dimensioni maggiori di quelle di altri formati immagine.

Esistono, infine, alcuni formati ISO basati sulla specifica TIFF 6.0 di Adobe (che è quella "ufficiale" del TIFF). Si tratta del formato ISO 12639, altrimenti noto come TIFF/IT, rivolto particolarmente al mondo del publishing e della stampa e dell'ISO 12234, altrimenti detto TIFF/EP, più orientato alla fotografia digitale.

3.2.2 JPEG / JPEG 2000

Sviluppato da	Joint Photographic Experts Group
Anno di rilascio prima versione	1986
Estensioni	.jpg, .jpeg, .jpe, .jif, .jfif, .jfi
Tipo MIME	image/jpeg
Formato aperto	Sì
Specifiche tecniche	pubblicate da Joint Photographic Experts Group

Standard	ISO/IEC 10918, ITU-T T.81, ITU-T T.83, ITU-T T.84, ITU-T T.86
Ultima versione	2009
Diffusione	Molto Alta
Collegamenti utili	http://www.jpeg.org/ www.iso.org

JPEG è il formato più utilizzato per la memorizzazione di fotografie. È inoltre il formato più comune su World Wide Web.

Lo stesso gruppo che ha ideato il JPG ha prodotto il JPEG 2000 (ISO/IEC 15444-1) che può utilizzare la compressione senza perdita di informazione. Il formato JPEG 2000 consente, inoltre, di associare metadati ad un'immagine. Nonostante queste caratteristiche la sua diffusione è tutt'oggi relativa.

3.3 Immagini Vettoriali

La grafica vettoriale è una tecnica utilizzata per descrivere un'immagine mediante un insieme di primitive geometriche che definiscono punti, linee, curve e poligoni ai quali possono essere attribuiti colori e anche sfumature.

I documenti realizzati attraverso la grafica vettoriale sono quelli utilizzati nella stesura degli elaborati tecnici, ad esempio progetti di edifici.

3.3.1 SVG

Sviluppato da	W3C (World Wide Web Consortium)
Anno di rilascio prima versione	2001
Estensioni	.svg, .svgz
Tipo MIME	image/svg + xml
Formato aperto	Sì
Derivato da	XML
Specifiche tecniche	Sviluppate da W3C
Standard	Raccomandazione del W3C
Ultima versione	SVG 2.0
Inserimento metadati	Sì
Collegamenti utili	http://www.w3.org/Graphics/SVG/

BOZZA

Diffusione	Alta
------------	------

Il formato SVG è un'estensione dell'XML, in grado di visualizzare oggetti di grafica vettoriale.

3.3.2 Shapefile

Sviluppato da	ESRI
Anno di rilascio prima versione	1990
Estensioni	obbligatori: .shp, .shx, .dbf opzionali: .sbn, .sbx, .fbn, .fbx, .ain, .aih, .prj, .shp.xml, .atx, .atx
Tipo MIME	application/octet-stream
Formato aperto	Sì
Specifiche tecniche	pubblicate da ESRI
Ultima versione	1998
Inserimento metadati	Sì
Collegamenti utili	http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf

Shapefile è un formato vettoriale per sistemi informativi geografici sviluppato dalla ESRI.

Il formato ha lo scopo di rendere interoperabili i sistemi ESRI con gli altri GIS. E' composto da un insieme di file di cui 3 obbligatori, con estensione .shp, .dbf, .shx.

3.4 Audio e Video

I formati audio/video, per la loro natura, non sembrano essere completamente compatibili con le caratteristiche indicate per i formati adatti alla conservazione.

Indipendenza dall'hardware, immodificabilità, assenza di software indesiderato, facilità di apporre la firma digitale non sono attribuibili a questi formati.

Tuttavia, vista la larga diffusione e il frequente uso che se ne fa anche da parte di alcuni settori della PA (Giustizia, ecc.), si è voluto metterne in evidenza almeno due tra quelli maggiormente supportati, sia per la standardizzazione e l'adeguamento delle specifiche, che per l'esistenza di numerosi software che si prestano al loro utilizzo.

3.4.1 MP3

Sviluppato da	MPEG (Moving Picture Experts Group)
Anno di rilascio prima versione	1992
Estensioni	.mp3
Tipo MIME	audio/mpeg
Formato aperto	No
Specifiche tecniche	pubblicate da MPEG
Standard	ISO/IEC 11172-3 (MPEG-1 Audio) ISO/IEC 13818-3 (MPEG-2 Audio)
Ultima versione	1995
Diffusione	Molto Alta
Collegamenti utili	http://www.mpeg.org www.iso.org http://mpeg.chiariglione.org/ http://mpeg.chiariglione.org/standards/mpeg-1/mpeg-1.htm http://mpeg.chiariglione.org/standards/mpeg-2/mpeg-2.htm

MP3 è un formato audio (codec audio) che usa un algoritmo di compressione di tipo *lossy*, sviluppato dal gruppo MPEG, in grado di ridurre notevolmente la quantità di dati richiesti per memorizzare un suono, rimanendo comunque una riproduzione accettabilmente fedele del file originale non compresso.

La qualità del suono di un file MP3 dipende dalla qualità della codifica e quindi del software utilizzato per questo scopo.

La sua diffusione è dovuta anche all'esistenza di numerosi software gratuiti che usano questo formato.

3.4.2 MPEG-4

Sviluppato da	MPEG (Moving Picture Experts Group)
---------------	-------------------------------------

BOZZA

Anno di rilascio prima versione	1998
Estensioni	.mp4
Tipo MIME	video/mp4
Formato aperto	Sì
Standard	da ISO/IEC 14496-1 a ISO/IEC 14496-28
Diffusione	Alta
Collegamenti utili	www.iso.org www.m4if.org/mpeg4/

L'MPEG-4 consiste in un gruppo di standard (ad oggi sono 28), utilizzato per applicazioni audio e video come la videotelefonia e la televisione digitale, la trasmissione di filmati via Web, e la memorizzazione di file audio-video su supporti CD-ROM.

3.5 Altri Formati

Per determinate tipologie di documenti informatici sono utilizzati specifici formati. In particolare in campo sanitario i formati più usati sono:

- DICOM (immagini che arrivano da strumenti diagnostici) anche se il DICOM non è solo un formato, ma definisce anche protocolli e altro;
- HL7 ed in particolare il CDA2 (Clinical Document Architecture) che contiene la sua stessa descrizione o rappresentazione.

Sul sito del Ministero della Salute si trovano le specifiche approvate di alcune tipologie di documenti quali le prescrizioni (<http://www.innovazionepa.gov.it/i-dipartimenti/digitalizzazione-e-innovazione-tecnologica/attivita/tse/il-tavolo-permanente-per-la-sanita-elettronica-delle-regioni-e-delle-province-autonome-tse-.aspx>).

4 Tabella dei formati per tipologie di uso

Tipo	Formato	Specifiche	Versione
Testi/Documenti	PDF/A	ISO 19005/1	1.7
Testi/Documenti	DOCX	ISO/IEC DIS 29500	
Testi/Documenti	ODT	ISO/IEC 26300:2006	
Immagini Raster	TIFF	pubblicate da Adobe	
Immagini Raster	JPEG	ISO/IEC 10918	

Immagini Vettoriali	DXF	pubblicate da Autodesk	
Immagini Vettoriali	SVG	raccomandazione del W3C	
Immagini Vettoriali	Shapefile	pubblicate da ESRI	
Audio	MP3	ISO/IEC 11172-3 (MPEG-1) ISO/IEC 13818-3 (MPEG-2)	
Video	MPEG-4	da ISO/IEC 14496-1 a ISO/IEC 14496-28	