

Documento n. 3 - Formati

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 sono stati scelti tra quelli che possono maggiormente garantire i principi dell'interoperabilità tra i sistemi di conservazione e in base alla normativa vigente riguardante specifiche tipologie documentali. 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 dell'Agenzia per l'Italia digitale.

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 usata per interpretare, leggere e modificare il file.

L'associazione del documento informatico al suo formato può avvenire, attraverso varie modalità, tra cui le più impiegate sono:

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. I metadati espliciti: l'indicazione "application/msword" inserita nei tipi MIME che indica un file testo realizzato con l'applicazione Word della Microsoft
3. il *magic number*: i primi byte presenti nella sequenza binaria del file, ad esempio 0xffd8 identifica i file immagine di tipo .jpeg

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, ...)
- Formati email (SMTP/MIME, ...)

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

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 .tif (usato dai fax), il .jpg, il .bmp.

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 proprietario per i file di tipo CAD, di cui non sono state rilasciate le specifiche;
- DXF, un formato simile al DWG, di cui sono state rilasciate le specifiche tecniche;
- Shapefile, un formato vettoriale proprietario per sistemi informativi geografici (GIS) con la caratteristica di essere interoperabile con i prodotti che usano i precedenti formati.
- SVG, un formato aperto, basato su XML, in grado di visualizzare oggetti di grafica vettoriale, non legato ad uno specifico prodotto.

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 e in particolare il CDA2 (Clinical Document Architecture) che contiene la sua stessa descrizione o rappresentazione.

L'informazione digitale è facilmente memorizzata, altrettanto facile è accedervi e riutilizzarla, modificarla e manipolarla, in altre parole, elaborarla e ottenere nuova informazione.

Questi formati, e i programmi che li gestiscono, che sono poi quelli che consentono e facilitano l'operatività giorno per giorno sul digitale, vanno valutati in funzione di alcune caratteristiche quali:

- La diffusione, ossia il numero di persone ed organizzazioni che li adotta
- La portabilità, ancor meglio se essa è indotta dall'impiego fedele di standard documentati e accessibili
- Le funzionalità che l'utente ha a disposizione per elaborare l'informazione e collegarla ad altre (ad esempio gestione di link)
- La capacità di gestire contemporaneamente un numero congruo (in funzione delle esigenze dell'utente) di formati
- La diffusione di visualizzatori che consentono una fruibilità delle informazioni in essi contenute indipendentemente dalla possibilità di rielaborarle.

Altre caratteristiche importanti sono la capacità di occupare il minor spazio possibile in fase di memorizzazione (a questo proposito vanno valutati, in funzione delle esigenze dell'utente, gli eventuali livelli di compressione utilizzabili) e la gestione del maggior numero possibile di metadati, compresi i riferimenti a chi ha eseguito modifiche o aggiunte.

È facilmente comprensibile come, nella fase di gestione del digitale, l'utente debba avere a disposizione la massima flessibilità possibile in termini di formati e funzionalità disponibili. Gli unici limiti sono quelli che un'organizzazione impone a se stessa quando, per esigenze di interscambio e interoperabilità, può determinare i formati e i relativi programmi di gestione che maggiormente soddisfano le contingenti esigenze operative.

Ai fini della formazione, gestione e conservazione, è necessario scegliere formati che possano garantire la leggibilità e la reperibilità del documento informatico nel suo ciclo di vita.

La scelta tra i formati dipende dalle caratteristiche proprie del formato e dei programmi che lo gestiscono.

Le caratteristiche di cui bisogna tener conto nella scelta sono:

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

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 questi ultimi garantiscano 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.

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.

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. Di fatto è indotta dall'impiego fedele di standard documentati e accessibili.

Per funzionalità si intende la possibilità da parte di un formato di essere gestito da prodotti informatici che prevedono una varietà di funzioni messe a disposizione dell'utente per la formazione e gestione del documento informatico.

Il supporto allo sviluppo è la modalità con cui si mettono a disposizione le risorse necessarie alla manutenzione e sviluppo del formato e i prodotti informatici che lo gestiscono (organismi preposti alla definizione di specifiche tecniche e standard, società, comunità di sviluppatori, ecc.).

La diffusione è l'estensione dell'impiego di uno specifico formato per la formazione e la gestione dei documenti informatici,

Questo elemento influisce sulla probabilità che esso venga supportato nel tempo, attraverso la disponibilità di più prodotti informatici idonei alla sua gestione e visualizzazione.

Inoltre nella scelta dei prodotti altre caratteristiche importanti sono la capacità di occupare il minor spazio possibile in fase di memorizzazione (a questo proposito vanno valutati, in funzione delle esigenze dell'utente, gli eventuali livelli di compressione utilizzabili) e la possibilità di gestire il maggior numero possibile di metadati, compresi i riferimenti a chi ha eseguito modifiche o aggiunte.

Per la scelta dei formati idonei alla formazione e gestione dei documenti informatici, sono da tenere in considerazione le caratteristiche indicate nei paragrafi precedenti. Ulteriori elementi da valutare sono l'efficienza in termini di occupazione di spazio fisico e la possibilità di gestire il maggior numero possibile di

metadati, compresi i riferimenti a modifiche o aggiunte intervenute sul documento. Le pubbliche amministrazioni indicano nel manuale di gestione i formati adottati per le diverse tipologie di documenti informatici motivandone le scelte effettuate; specificano altresì i casi eccezionali in cui non è possibile adottare i formati in elenco motivandone le ragioni.

La scelta dei formati idonei alla conservazione, oltre al soddisfacimento delle caratteristiche suddette, deve essere strumentale a che il documento assuma le caratteristiche di immodificabilità e di staticità previste dalle regole tecniche. Per quanto fin qui considerato, è opportuno privilegiare i formati che siano standard internazionali (de jure e de facto) o, quando necessario, formati proprietari le cui specifiche tecniche siano pubbliche, dandone opportuna evidenza nel manuale di conservazione dei documenti informatici. Ulteriore elemento di valutazione nella scelta del formato è il tempo di conservazione previsto dalla normativa per le singole tipologie di documenti informatici. I formati per la conservazione adottati per le diverse tipologie di documenti informatici devono essere indicati nel manuale di conservazione motivandone le scelte effettuate; sono altresì specificati i casi eccezionali in cui non è possibile adottare i formati in elenco motivandone le ragioni.