PROGETTO DI CONSERVAZIONE DIGITALE
A LUNGO TERMINE DEI MANOSCRITTI
DELLA BIBLIOTECA APOSTOLICA VATICANA

Digitalizzare per conservare e divulgare

Papiro Bodmer XIV-XV (P_{75}), f. 1B2$_v$
(Luca 11,1-13; il *Padre Nostro* si legge nelle righe 7-13)

Papiro Bodmer XIV-XV (P_{75}), f. 2A8$_r$
(Luca 24,51-53 e Giovanni 1,1-16)

Quale migliore testimonianza di due antichissime pagine del vangelo di Luca e di Giovanni per lanciare un messaggio forte sul diritto al salvataggio e alla divulgazione di queste importanti opere che conservano la nostra storia!

La Biblioteca Apostolica Vaticana, dalla sua fondazione più di cinque secoli fa, conserva, protegge e restaura il proprio immenso e prezioso patrimonio librario: un vero tesoro dell'umanità custodito a servizio di tutti.

Tuttavia, proprio il tempo è il nemico peggiore per la conservazione di questo grande patrimonio culturale: microorganismi, umidità e altri fattori, anche il semplice contatto con la pelle umana, giorno dopo giorno, nonostante l'impegno profuso, contribuiscono a deteriorare questi delicati documenti della nostra storia.

Per preservare i suoi manoscritti, la Biblioteca Vaticana ha avviato fin dal 2007 uno studio di fattibilità per comprendere come le tecnologie informatiche potessero aiutare a risolvere il problema di un'adeguata conservazione, giungendo alla conclusione che la soluzione fosse quella di digitalizzare il suo intero patrimonio manoscritto.

Mentre il progetto di digitalizzazione prendeva lentamente forma, la Biblioteca avviava un'articolata analisi dei molteplici aspetti implicati nel processo di digitalizzazione, allestendo una procedura di *TestBed* attraverso la quale realizzare una microarea in cui fossero presenti tutte le funzionalità e le modalità del progetto globale.

Durante la fase di *TestBed* La Biblioteca si pose l'obiettivo di stabilire procedure e formati. Se si pensa alla rapida obsolescenza di quasi tutto quello che ha a che fare con l'informatica, sia hardware che software, si capisce come il lavoro di progettazione e poi di realizzazione non sia per niente facile. Una volta digitalizzati i manoscritti, le loro immagini saranno messe a disposizione sia degli studiosi, che potranno lavorare su copie digitali ad altissima risoluzione e con colori assolutamente fedeli agli originali, sia del grande pubblico, in quest'ultimo caso a risoluzioni inferiori, più gestibili tramite appositi siti internet. In questo modo tutti nel mondo avranno la possibilità di ammirare la bellezza, studiare i contenuti e capire l'importanza di questi manoscritti e libri antichi.

C'è da considerare inoltre che avere un'intera biblioteca di manoscritti e libri antichi scansionati in formato digitale ad altissima risoluzione – cosa mai avvenuta finora – potrà permettere, grazie alla moderna tecnologia, studi incrociati, comparazione tra immagini, riconoscimento di scritture e tutta una serie di operazioni che finora richiedevano molto tempo e che soprattutto necessitavano della presenza fisica del manoscritto antico nelle mani dello studioso, con tutte le conseguenze negative in termini di usura e "stress" dell'oggetto.

I principi fondamentali sui quali abbiamo fondato le linee guida del progetto di conservazione digitale a lungo termine, sono i seguenti: il formato di conservazione, l'attenta analisi dei principi di controllo relativi all'obsolescenza tecnologica dei sistemi usati per lo *storage* e l'analisi qualitativa delle immagini, inclusa la costante taratura sul bilanciamento dei colori tra monitor e apparati di acquisizione.

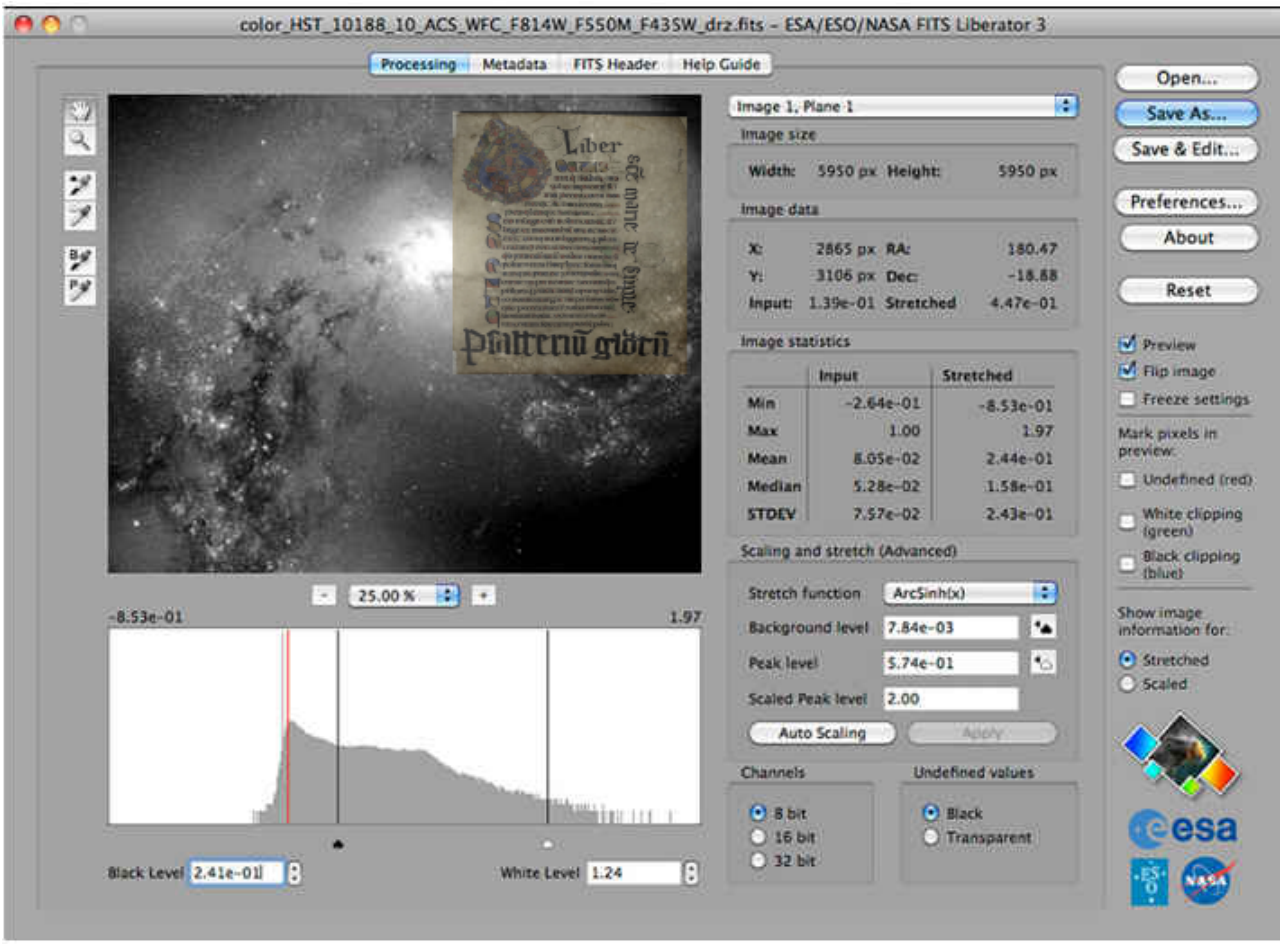
Dopo attenti studi comparativi la scelta si è focalizzata sul formato FITS, che a nostro giudizio possiede la gran parte dei requisiti richiesti:

- È un formato progettato dalla NASA negli anni Settanta e reso di *public domain* con distribuzione gratuita dei codici di *source*. Il suo aggiornamento è garantito da tutta la comunità scientifica di astrofisica e fisica spaziale mondiale ogni 6 mesi attraverso l'organizzazione IAU, l'ente che governa tutta la comunità scientifica afferente al FITS.
- Questo formato riesce a documentare in modo esaustivo il contenuto dell'immagine salvata con un gran numero di informazioni aggiuntive.
- Con questo formato sarà possibile fin d'ora gestire immagini di grandi dimensioni superiori a 4 giga byte.
- È un formato estremamente flessibile.
- Ha dimostrato la capacità di seguire l'evoluzione tecnologica del settore (per es. il passaggio praticamente indolore dai sistemi a 32 bit a quelli a 64 bit).
- Ha un'esperienza consolidata da oltre 40 anni di attività e una comunità scientifica che lo sostiene e aggiorna ad intervalli regolari di tempo.

- Possiede molte keywords che è possibile usare per immagazzinare informazioni similari provenienti da altri formati.
- È anche pronto per l'acquisizione in 3D o a livelli superiori.

La comparazione con il più noto TIFF ha fatto emergere molte lacune su quest'ultimo formato:

- Il TIFF infatti è un formato proprietario della soc. ADOBE e non rilascia liberatorie sulla totale gratuità neanche per grandi progetti.
- Il TIFF è stato progettato nel 1992 ma il suo ultimo aggiornamento risale al 1998.
- Il TIFF è un formato a 32 bit per cui i file generati con questo formato non possono superare i 4 giga byte.
- Il TIFF non è progettato per la terza dimensione.
- Il TIFF ha una gestione propietaria delle keyword dei metadati.



Naturalmente ci siamo resi conto che anche il FITS aveva a dei "fattori negativi":

- Il fatto che non fosse nato in maniera esplicita per la gestione di immagini fotografiche fa sì che non siano molti i programmi di uso comune in grado di visualizzare e gestire immagini FITS; per rimediare a questo aspetto, si è scelto di salvare le immagini usando il profilo colore denominato sRGB in modo da renderle subito visualizzabili e gestibili con alcuni software molto diffusi nel settore, come Adobe Photoshop o GIMP.
- Alcune informazioni tipiche di quel settore non sono immediatamente disponibili, per es. la risoluzione dell'immagine (pixel/unità di misura), necessaria per capire la qualità della scansione. Infatti lavorando su manoscritti di dimensioni molto variabili il numero di pixel totali dell'immagine può non essere sufficiente a capire questo importante parametro. Un'immagine di 8192x6286 punti ha una qualità diversa se queste dimensioni sono relative a un piccolo foglio oppure a una grande pergamena.
- Non memorizza in maniera nativa l'ICC profile, vale a dire quelle informazioni e caratteristiche tipiche del dispositivo di scansione che permettono successivamente una riproduzione assolutamente fedele dei colori originali.
- Per sanare queste incoerenze in collaborazione con la Facoltà di Astrofisica e Fisica Spaziale di Roma abbiamo elaborato un *asset* delle *keyword* nell'*History file* del FITS che sia in grado, nei processi di riconversione, di ereditare correttamente tutte le informazioni derivanti da acquisizioni in formato TIFF.

Dopo aver stabilito che formato usare, si è iniziato a definire nel dettaglio il processo di salvataggio digitale dei manoscritti, che inizia con la scansione dei manoscritti mediante particolari scanner piani e macchine fotografiche, che prima di tutto preservano il manoscritto durante le

operazioni di acquisizione delle immagini ad altissima risoluzione, evitando per esempio di forzare le piegature dei libri per facilitare la scansione.

Infatti, poiché spesso con i libri antichi non è possibile aprire completamente le pagine, rimane una curvatura delle stesse più o meno accentuata; è quindi stato sviluppato un software *ad hoc* per misurare questa curvatura, elaborare l'immagine e salvarla come se fosse stata acquisita in modo perfettamente piano, mantenendo inalterate proporzioni e distanze.

Successivamente l'immagine viene salvata in formato TIFF, formato standard di uscita per la totalità dei *device* del settore e tuttora il più usato nel mondo della fotografia e nella visualizzazione nel mondo dei beni culturali.

Il formato TIFF normalmente usato contiene campi essenzialmente con informazioni di natura "fotografica" che in molti casi possono avere una corrispondenza nel formato FITS.

TABELLA DI CORRISPONDENZA TRA TAG TIFF E TAG FITS NELL'AMBITO DEL PROGETTO DI DIGITALIZZAZIONE DELLA BAV

Nota: le keyword in verde sono quelle per cui ancora non è stato determinato un corrispondente in FITS

TIFF					FITS		
Dec	Hex	NAME	VALUE	Short description	KEYWORD	Type	Description
256	0100	ImageWidth	Short Long	The number of columns in the image, i.e., the number of pixels per row.	NAXIS1 (Standard)	Integer	
257	0101	ImageLength	Short Long	The number of rows of pixels in the image.	NAXIS2 (Standard)	Integer	
258	0102	BitsPerSample	Short	Number of bits per component.	BITPIX (Standard)	Integer	In tiff il valore è una terna 8, 8, 8, in FITS un solo integer 8 bit (profondità)
262	0106	PhotometricInterpretation	Short	The color space of the image data.	COLORMAP	Integer	Il profilo colore sRGB è lo standard utilizzato per i nostri FITS.
272	0110	Model	ASCII	The scanner model name or number.	INSTRUME (Standard)	String	Incorpora le informazioni contenute in MODEL e MAKE. Es. HP Scanjet 4400.
274	0112	Orientation	Short	The orientation of the image with respect to	ORIENTAT	Integer	Rappresenta l'inclinazione dell'asse Y in senso orario

				the rows and columns.			rispetto al Nord. Ad es. 0 = Verticale 90 = Orizzontale
277	0115	SamplesPerPixel	Short	The number of components per pixel.	NAXIS (Standard)	Integer	Numero di componenti Ad es. 3 (esempio di immagine RGB)
282	011A	XResolution	Rational	The number of pixels per ResolutionUnit in the ImageWidth direction.	XRES	Float	Risoluzione dell'immagine sull'asse X.
283	011B	YResolution	Rational	The number of pixels per ResolutionUnit in the ImageLength direction.	YRES	Float	Risoluzione dell'immagine sull'asse Y.
296	0128	ResolutionUnit	Short	The unit of measurement for XResolution and YResolution.	RESUNIT	Integer	Unità di misura per la risoluzione. Ad es 2 Per indicare Inch
305	0131	Software	ASCII	Name and version number of the software package(s) used to create the image.	PROGRAM (Common Used)	String	Software utilizzato per la creazione dell'immagine.
306	0132	DateTime	ASCII	Date and time of image creation.	DATE (Standard)	String	Data e Ora dell'acquisizione in formato internazionale yyyy-mm-ddThh:mm:ss
315	013B	Artist	ASCII	Person who created the image.	AUTHOR (Standard)	String	Autore che ha creato l'immagine.
33432	8298	Copyright	ASCII	Copyright notice.	ORIGIN (Standard)	String	Copyright sull'immagine.
34675	8773	ICC Profile	Undefined	ICC profile data.	-	String	Se necessario utilizzare più tag TAGFROM. Dato che ICCProfile può essere più lunga di 80 caratteri prevediamo di distribuire l'informazione in più righe e in esadecimale. Ad es. TAGFROM TIFF;ICCProfile;34675 = xxxx TAGFROM TIFF;ICCProfile;34675 = yyyy etc etc
-	-	-	-	-	REFERENC	String	Indicazione relativa alla posizione del file XML collegato al file FITS.

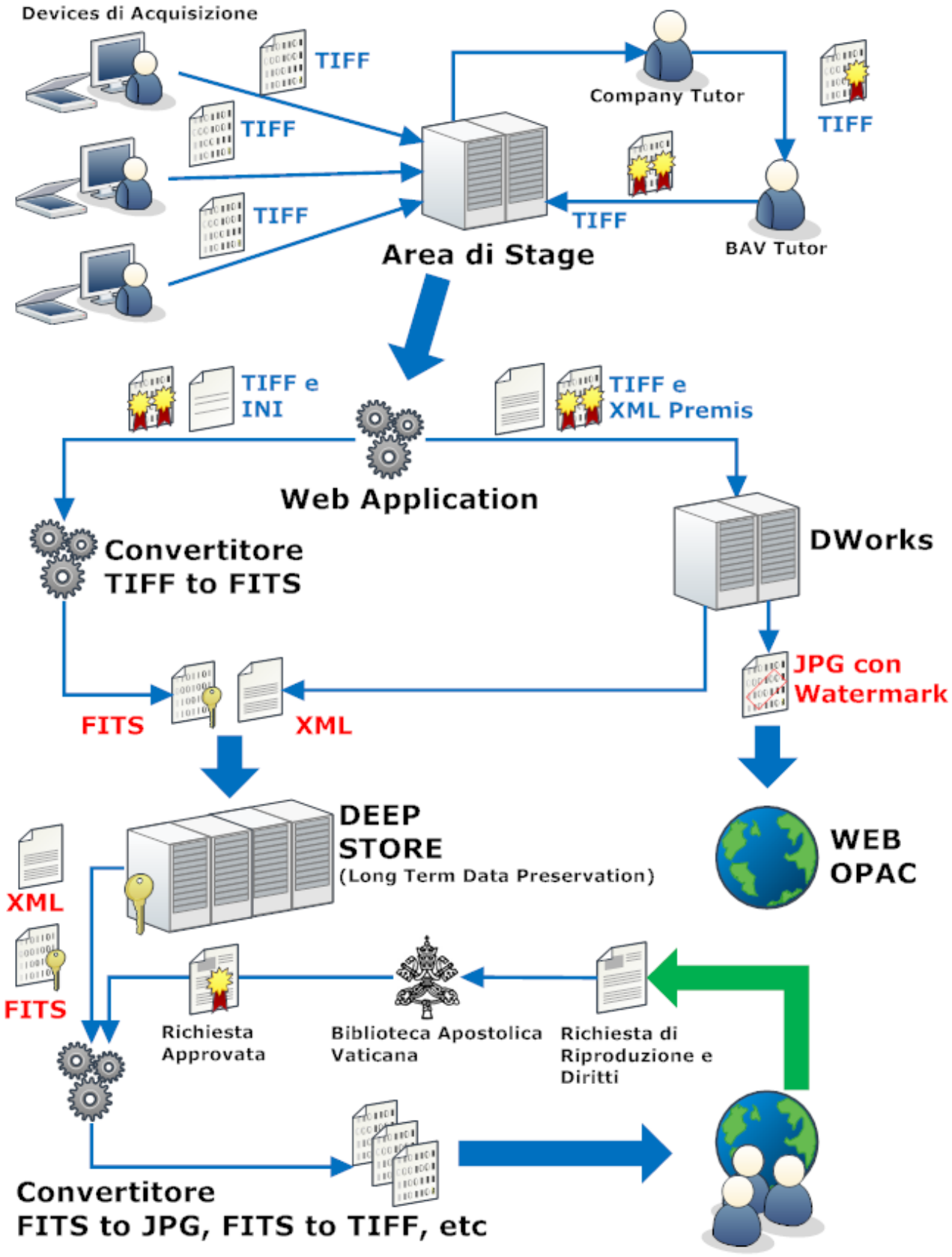
Purtroppo per alcune di queste informazioni non abbiamo riscontrato un equivalente nel FITS. Per questo motivo abbiamo proposto allo IAU (l'organizzazione mondiale che governa l'aggiornamento del formato FITS nel mondo) la creazione di alcune nuove *keyword* che permetteranno di trasformare in FITS i file provenienti da altri formati grafici conservando tutte le informazioni ritenute utili.

Il *workflow* ha origine dai *devices* di acquisizione delle immagini. I devices producono immagini TIFF ad altissima definizione e file XML *premis*. Questi file vengono immagazzinati in attesa di elaborazione nell'area di stage. Lo spazio attualmente disponibile nell'area di *storage* ci permette di conservare i file prodotti in 6 mesi. Questa area, denominata "di *stage*", è strutturata secondo architetture EMC² su macchine ISILON. Dopo questa fase i file TIFF vengono convertiti in FITS e destinati allo *storage* permanente di tipo ATMOS, sempre in tecnologia EMC².

Un applicativo *web-based* si occupa di gestire, tramite un'apposita interfaccia, e di monitorare tutti i file acquisiti; se essi sono conformi alle attese, l'applicativo li manda come *imput* ai vari processi, qui di seguito elencati:

- L'interfaccia *web* è lo strumento mediante il quale due *tutor* (esperti formati del settore) distinti controllano e validano l'immagine acquisita: essi sono uno di una società esterna e uno della Biblioteca Vaticana. Se le immagini del volume digitalizzato sono validate e approvate, l'applicativo *web* si occuperà di:

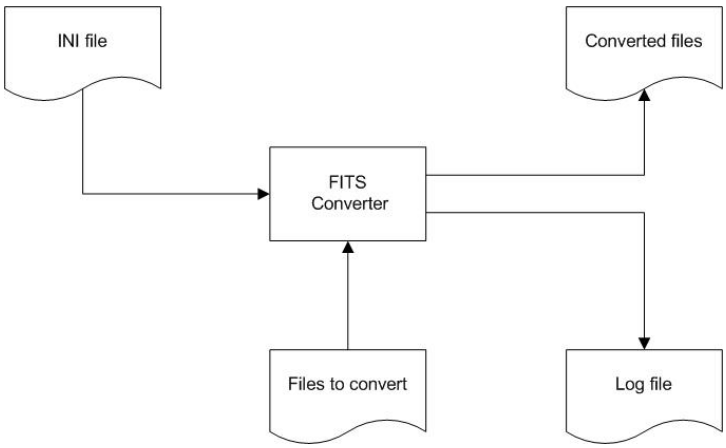
1) avviare il processo di trasferimento nell'area di *stage* dei TIFF;
2) avviare il processo di conversione in JPG a bassa risoluzione, timbratura dei file, generazione e indicizzazione del file XML partendo dal XML *premis* file e infine pubblicazione su internet per la visualizzazione delle immagini agli utenti del *world wide web*;
3) dopo sei mesi convertire i file TIFF presenti nell'area di stage in FITS per il processo di lunga conservazione digitale su sistemi ATMOC EMC².



In dettaglio, il processo di conversione da TIFF a FITS prevede la realizzazione di una componente *software* in *Java*, che permette non solo di

convertire i file TIFF in formato FITS, ma anche di far migrare, all'interno del FITS, alcuni importanti TAG presenti nell'immagine sorgente e, contestualmente, altri presenti in un file INI generato dall'applicativo *web-based* che gestisce e monitorizza le immagini scansionate del patrimonio librario memorizzate temporaneamente nell'area *stage*.

Qui di seguito viene riportato il flusso base di conversione:



Per poter eseguire la conversione è necessaria la presenza del file INI e, ovviamente, il gruppo di file immagine interessati alla conversione. Il convertitore genera i file FITS in un percorso specifico (dichiarato sul file INI) e contestualmente genera anche un *log* delle operazioni svolte. In questo modo la *web application* può rendersi conto, analizzando il *log*, che tutte le conversioni sono state eseguite correttamente o prendere eventuali contromisure per una mancata conversione.

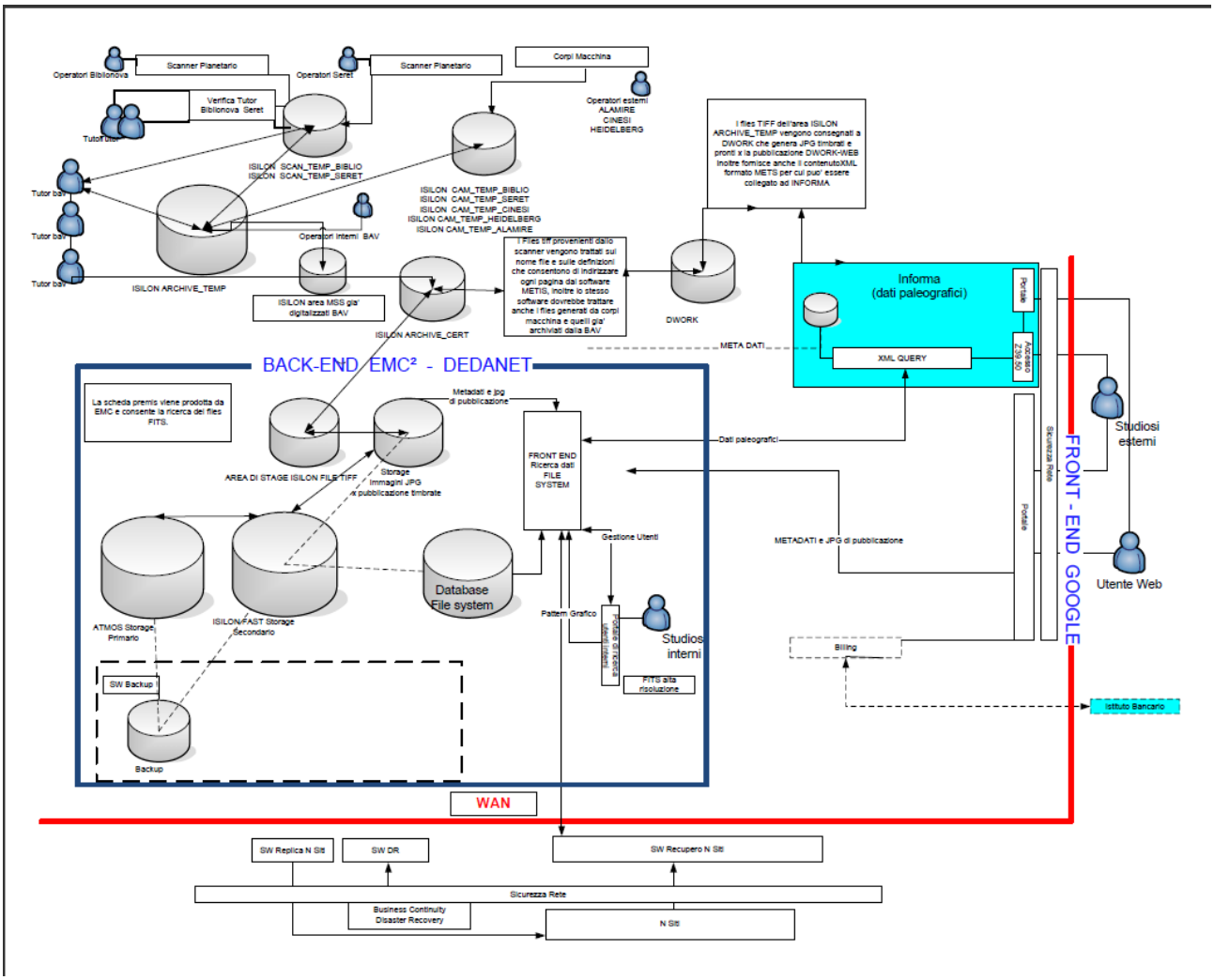
Il secondo processo avviato dall'applicazione *web* prevede la conversione dei file TIFF acquisiti in formato JPG a bassa risoluzione. I file JPG, prima di essere pubblicati sul *web server* che li renderà visibili al mondo, verranno timbrati con il *copyright* della Biblioteca Apostolica Vaticana**.** La presenza del *copyright* è volta essenzialmente a evitare il più possibile un uso non autorizzato dei file.

Contestualmente, rielaborando le informazioni presenti nell'XML *premis*, verrà generato un file XML per ogni immagine, che conterrà sia

informazioni bibliografiche dell'immagine scansionata sia informazioni sulle modalità e tecnologie utilizzate durante il processo di acquisizione.

Il file XML verrà indicizzato in una base di dati, e le informazioni archiviate saranno parte integrante, insieme alle immagini JPG, dell'*Open Public Access Catalog* utilizzabile dagli utenti del *world wide web*.

Schema del flusso di lavoro nel processo di conservazione digitale a lungo termine:



Per quanto concerne gli altri due punti espressi nelle linee guida – l'obsolescenza del *data center* e l'obsolescenza delle apparecchiature di conservazione o *storage* – si tratta di aspetti strettamente correlati alla manutenzione di questi stessi apparati, con l'aggiunta di una ulteriore attenzione per tutti gli apparati di acquisizione digitale.

Deve infatti essere chiaro che un severo controllo periodico di queste apparecchiature elettroniche di acquisizione è indispensabile per verificare la resa dagli scanner planetari di qualsiasi marca o modello utilizzati in progetti di digitalizzazione di materiale bibliografico antico e raro.

I controlli qualitativi adottati si ispirano allo standard denominato UTT Universal Test Target (http://www.universaltesttarget.com/index.php) rilasciato da

- National Library of the Netherlands (KB)
- Image Engineering Dietmar Wueller (IE)
- Fachverband für Multimediale Informationsverarbeitung

come standard aperto per la verifica delle immagini digitalizzate da qualsiasi tipo di scanner e sistema di ripresa digitale.

UTT ha definito un Test Target Universale da utilizzare per la verifica della qualità delle immagini prodotte (digitalizzate).

Con un unico *test target* possono quindi essere eseguite, attraverso una scansione periodica (in relazione alle procedure adottate per il progetto di digitalizzazione), diverse misurazioni in conformità ai più diffusi standard ISO.

Il test target si compone, in relazione alla dimensione del piano di scansione dello scanner, di una o più immagini A3, ed è rilasciato su supporto indeformabile, con i relativi dati di referenza personalizzati per ogni singola *test chart*.

Le misurazioni che possono essere realizzate, utilizzando il *test target* UTT sono:

- MTF (Modulation Transfer Function) in conformità allo standard ISO 16607;
- Color Reproduction (Spazi Colore supportati: CIE 1976, CIE 1994, CIE 200, etc);
- Livello del Noise in conformità allo standard ISO 12233;
- OECF (Opto Electronic Conversion Function) in conformità allo standard ISO 14523;

- Homogeneity (omogeneità della illuminazione) sull'intero piano di scansione;
- Distorsion (risoluzione sull'asse X, risoluzione sull'asse Y e analisi delle aberrazioni geometriche).

Come accennavo, le analisi del *test target* possono anche essere effettuate eseguendo una scansione della stessa immagine e analizzando il file *raw tif* che ne deriva con una applicazione proprietaria denominata OS QM-Tool rilasciato da Zeutschel. Questo *software* OS QM-Tool effettua l'analisi della scansione in pochi secondi e restituisce un file di *log* che riporta tutte le misurazioni effettuate. Esso permette, inoltre, di impostare, per ciascun parametro da misurare, dei valori assoluti, che potranno essere assunti quale soglia di *warning* e/o di errore, qualora l'analisi rilevi un valore superiore.

La misurazione delle specifiche tecniche dei sistemi di digitalizzazione è uno strumento utile a determinare la qualità delle immagini riprodotte. È quindi, buona norma eseguire, all'inizio di ciascun progetto di digitalizzazione, una verifica di qualità che permetta di fissare lo *standard* di riproduzione e che in seguito potrà essere utilizzato come *standard* di riferimento per monitorare con una frequenza costante (es. ogni 2.000 scansioni) l'omogeneità di riproduzione digitale durante l'intera durata del progetto.

Ovviamente lo strumento citato è suggerito fra i molti esistenti sul mercato, altrettanto validi.

Con la speranza di aver lanciato un sassolino nel mare della conservazione digitale, vi ringrazio per la vostra attenzione.

Luciano Ammenti