

La conservazione a lungo termine delle risorse digitali: cosa si può fare oggi

Firenze, 18 giugno 2008

Test iniziale

- ✓ dc
- ✓ RSS
- ✓ RDA
- ✓ XML
- ✓ md5

Il problema

- ✓ DP non solo una questione tecnologica (evitare la corruzione dei bit o l'obsolescenza degli strumenti informatici)
- ✓ Ma anche:
 - ✓ sostenibilità economica,
 - ✓ l'affidabilità dei sistemi di deposito,
 - ✓ i criteri di selezione di quello che *vale la pena* conservare,
 - ✓ il quadro legislativo,
 - ✓ i ruoli e le responsabilità istituzionali
 - ✓ ecc.

Di che cosa parliamo - 1

- ✓ file
- ✓ gestiti sia come sequenze di bit *tout court*
- ✓ ma anche come sequenze di bit che rispondono a un determinato formato
- ✓ (possono essere viste come *insieme di dati e metadati* e si può parlare di *metadati interni*);

Di che cosa parliamo - 2

- ✓ risorse digitali (composte da uno o più file) . Possiamo definire risorsa come "qualsiasi cosa che sia identificabile", ad esempio "un documento elettronico, una immagine, un servizio - che tempo fa oggi a Los Angeles - possono essere considerati esempi di risorse": naturalmente "non tutte le risorse sono reperibili in rete: ad esempio anche gli esseri umani, le società e i libri in una biblioteca possono essere visti come risorse"

Di che cosa parliamo - 2

- ✓ metadati *associati* alla risorsa (metadati esterni alla risorsa). Si tratta di informazioni che ci permettono di identificare e di gestire una determinata risorsa per determinate funzioni.
 - ✓ Ad esempio informazioni sui diritti di accesso alla risorsa sono funzionali alla fruibilità, mentre le informazioni sull'impronta dei file che compongono la risorsa sono essenziali sia per stabilire la vitalità dei singoli file, sia per stabilire l'autenticità della risorsa che stiamo esaminando

OAIS (Metadati e risorsa) -1

- ✓ Il modello OAIS vede la risorsa digitale (*Content information* nella terminologia OAIS) come inseparabilmente composta da:
 - ✓ *A.1 Content data object* che consiste in una sequenza (stream) di bit o in un set di sequenze di bit
 - ✓ e da *A.2 Representation information*

OAIS (Metadati e risorsa) -2

- ✓ *A2 Representation information* =
 - ✓ *metadati* che traducono il *Content data object* in conoscenza accessibile - per esempio processabile da un computer - e dotata di significato - per esempio comprensibile ad un essere umano).
 - ✓ Come esempio di *Representation information* si può pensare alla codifica di un documento eseguita tramite il programma Microsoft Word

OAIS (Metadati e risorsa) -3

- ✓ *B. Descriptive information:* i famosi *metadati descrittivi*
- ✓ *C. Packaging information:* danno informazioni relative al tipo di rapporto tra risorsa e supporti che la veicolano (per esempio in che server in che directory ecc. si trova quel determinato file) e come sono collegati i dati ai metadati;
- ✓ *D. Preservation description information:* veicolano informazioni che hanno come obiettivo primario la conservazione nel tempo della risorsa

OAIS (Metadati e risorsa) - 4

Risorsa

Metadati "esterni"

A1 = 01010111 ...

B = descrizione

A2 = metadati "chiave"

Representation inf.

(anche formato file)

C = risorsa /
supporto; dati /
metadati

D = conservazione

risorsa / supporto - 1

Libro

CD

Immagine
jpeg

Supporto/Pubblicazione

- Indipendenza dal supporto
- Le "dipendenze" del digitale

risorsa / supporto - 2

- ✓ "la rappresentazione informatica di atti, fatti o dati giuridicamente rilevanti "[DPR 10 novembre 1997, n. 513]
- ✓ "per documento informatico si intende qualunque supporto informatico contenente dati o informazioni aventi efficacia probatoria o programmi specificamente destinati ad elaborarli" [Articolo 491-bis del codice penale, introdotto dalla legge 23 dicembre 1993, n. 547]

risorsa / supporto - 3

- ✓ rappresentazione informatica = “una sequenza di bit, che, elaborata da un sistema informatico, può essere resa visibile su uno schermo, stampata sulla carta o inviata a distanza”
- ✓ “cambiamento radicale: il documento da
 - ✓ res signata, cioè di una cosa che riporta dei segni, delle informazioni;
 - ✓ a funzione autonoma rispetto alla sua (eventuale) fissazione su un supporto materiale”

[Cammarata - Maccarone 2000]

Libro, cd, immagine jpg

- ✓ Vitalità = Viability (catene di 0 e di 1 intatte)
- ✓ Rappresentabilità (Traducibilità) = Renderability (attraverso un elaboratore)
- ✓ Fruibilità = Understandability (da parte di esseri umani)

Risorsa /supporto - 4

	Libro	CD	Immagine jpeg
Vitalità	Forte dip. Supporto	Non essenziale dip. supporto	Nessuna dip. supporto
Rappresentabilità (Traducibilità)	Basse dip. tecnologiche	Alte dip. tecnologiche	Alte dip. tecnologiche
Fruibilità	Dip dalla catena di informazioni contestuali (B-C-D)		

Conservazione come

- ✓ Servizio pubblico in senso oggettivo (non soggettivo)
- ✓ Fornito da depositi digitali accreditati (es. TRAC)
- ✓ Allo scopo di assicurare per le risorse digitali depositate:
 - ✓ Vitalità
 - ✓ Traducibilità
 - ✓ Autenticità
 - ✓ Fruibilità

Vitalità

✓ Viability

- ✓ le sequenze di bit che compongono i file sono intatte
- ✓ Condizione necessaria ma non sufficiente
- ✓ Non sempre “utile”: formati protetti, DRM ecc

Traducibilità

- ✓ Rappresentabilità da parte di un elaboratore
- ✓ Un determinato hardware e un determinato software sono in grado di gestire le risorse digitali depositate (= riprodurre le funzionalità progettate dall'autore)

Autenticità (1)

- ✓ Documentazione di
 - ✓ Identità (quello che lo distingue dagli altri documenti)
 - ✓ Integrità (il messaggio che il documento intendeva comunicare è inalterato)

Autenticità (2)

- ✓ Differenza tra
 - ✓ Autenticazione (firma digitale, sigillo)
 - ✓ Autenticità (proprietà del documento)

Fruibilità

- ✓ Da parte della “ comunità di riferimento” (quella che ha conferito l’incarico)
- ✓ Rendere possibili servizi quali:
 - ✓ Trovare
 - ✓ Identificare
 - ✓ Selezionare
 - ✓ Ottenere

Altre definizioni

- ✓ Conservazione del digitale come
 - ✓ Mezzo per rendere possibile l'accesso nel tempo
 - ✓ Comunicazione (Conversazione) differita

La conservazione quotidiana

- ✓ Personale
- ✓ Istituzionale

Personale - 1

Nascondere hard disk ...

“Dai tempi di Poggiolini fino a quelli attuali di Stefano Ricucci

l'idea che i propri dati "scottanti" possano prendere strade "immateriali" (che ne so un banale account anonimo su Gmail) che ne rendano sicuro l'occultamento fatica a farsi strada.

Nascondere due hard disk in un divano continua ad avere quel sapore amaro di rifiuto della tecnologia che e' oggi una delle piccole (ho scritto "piccole") tragedie di questo paese.”
(Massimo Mantellini)

Personale - 2

- ✓ La nostra eredità digitale:
 - ✓ Su web
 - ✓ Nei file dei computer (casa ufficio etc)
 - ✓ Nei CD e nelle chiavi USB ... dimenticati

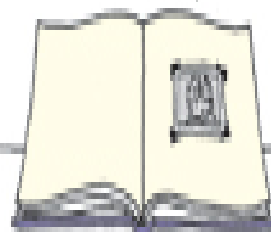
Personale - 3

- ✓ Non si può dimenticare in soffitta
- ✓ Non esiste il CD eterno e anche se esistesse non sarebbe di alcuna utilità
- ✓ Puntare su:
 - ✓ Molteplicità
 - ✓ Storage on line
 - ✓ Sistemi personali di gestione delle password
 - ✓ Formati standard

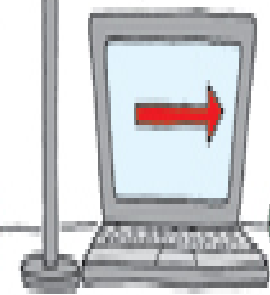


THEN

ONE PHOTO OF ONE EVENT
IN ONE ALBUM IN ONE ATTIC.



NOW

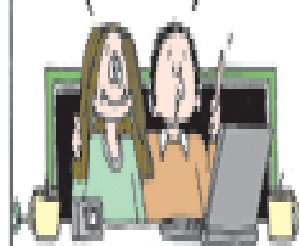


ALL OUR PHOTOS
ARE SAFELY STORED
ON THE COMPUTER!



WHAT IF
THE
COMPUTER
CRASHES?

THEY'RE
BACKED UP TO
AN INTERNET
STORAGE SITE!



WHAT IF
THE SITE
GOES
DOWN?

THEY'RE BACKED
UP TO AN
EXTERNAL, PORTA-
BLE HARD DRIVE!



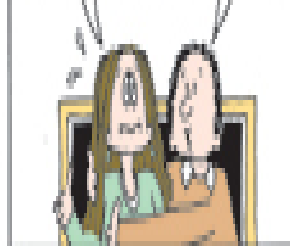
WHAT IF
THE
EXTERNAL
DRIVE FAILS?

THEY'RE
BACKED UP
TO CD!



WHAT
IF THE
CDs
DETERI-
ORATE
???

THEY'RE BACKED
UP TO A TAPE
DRIVE STORED
IN A CLIMATE-
CONTROLLED
VAULT !!



WHAT IF THE TECHNOLOGY
CHANGES AND TAPE IS NO
LONGER READABLE BY THE
EQUIPMENT OF THE DAY??



OLD OBSESSION: FINDING OURSELVES.
NEW OBSESSION: PRESERVING OURSELVES.

AT LEAST WE'LL
ALWAYS HAVE
THE PRINTS.

WHAT IF THEY WEREN'T PRINTED ON
ACID-FREE, ARCHIVAL PAPER WITH
NON-FADING, WATER-RESISTANT INK??



© 2008 Cathy Curcote. All rights reserved.

La conservazione quotidiana - Istituzionale

- ✓ con il digitale
 - ✓ occorre che l'istituzione dichiari esplicitamente una politica di conservazione
 - ✓ Non è ammessa la discontinuità

(Giordano 2007)

Istituzionale - 2

- ✓ In una politica di conservazione rientra:
 - ✓ L'uso di software aperti
 - ✓ L'uso di formati aperti

Istituzionale - formati di File

- ✓ Representation information (anche "formato" del file)
 - ✓ Libere (es. html, xml ecc)
 - ✓ Proprietarie
 - ✓ Pubbliche (es pdf)
 - ✓ Riservate (es doc)
 - ✓ PDF/A

Istituzionale problemi da affrontare

- ✓ Rapida obsolescenza tecnologica
- ✓ Errori umani
- ✓ Disastri naturali
- ✓ Attacchi esterni /interni
- ✓ Problemi economici e organizzativi
- ✓ Produciamo una quantità enorme di digitale, vogliamo accedere a tutto velocemente e più a lungo
- ✓ (Gladney 2007)

Strategie per la traducibilità

- ✓ Migrazione
- ✓ Emulazione
- ✓ Normalizzazione a XML

Migrazione

- ✓ Le informazioni vengono scaricate dal vecchio sistema e caricate sul nuovo
- ✓ E' da tempo praticata con successo
 - ✓ Per esempio con i database (fatti per rendere i dati indipendenti dalle applicazioni -> una anagrafe o una biblioteca cambiano sistema di gestione)
- ✓ Può comportare la perdita di informazioni

Emulazione

- ✓ E' basata su due assunti:
 - ✓ Il funzionamento dell'hardware è di solito documentato in maniera aperta per invogliare i produttori di software a farne uso
 - ✓ In futuro esisteranno elaboratori (qualcosa in grado di elaborare le catene di 0 e di 1)

Emulazione - 2

- ✓ Per archiviare una presentazione (ppt) occorre archiviare:
 - ✓ Le specifiche di funzionamento dell'hw (Intel xxx)
 - ✓ Il sistema operativo Windows XP
 - ✓ Il programma Powerpoint
 - ✓ La presentazione

Emulazione - 3

- ✓ In futuro (tra x anni) sarà possibile:
 - ✓ Costruire un macchina virtuale in grado di emulare l'hardware del 2002
 - ✓ Caricare su quella macchina virtuale
 - ✓ Windows XP
 - ✓ Powerpoint
 - ✓ La presentazione

Emulazione 4

- ✓ Vmware
- ✓ Virtual Box
- ✓ Mame
- ✓ Universal Virtual C
- ✓ Un esempio



Normalizzazione (XML)

- ✓ Decisione di un archivio in fase di ingest (accettare solo pochi formati, documentati - XML come formato che si auto-documenta)
- ✓ Il problema di due standard XML per i documenti

Normalizzazione (XML) - ma quale?

- ✓ questo?

- ✓ oppure questo- voluto da Microsoft?

Il commento di Tim Bray (padre di XML)
"What Microsoft really wanted was that ISO stamp of approval to use as a marketing tool.

- ✓ And just like your mother told you, when they get what they want and have their way with you, they're probably not gonna call you in the morning".

Livelli MD

- ✓ **Vitalità** (... le sequenze di bit - file - sono integre e leggibili - come sequenze - da un elaboratore dal primo all'ultimo bit)
- ✓ **Traducibilità** (le sequenze di bit sono interpretate da un elaboratore come insieme di dati e metadati)
- ✓ **Autenticità** (vi sono sufficienti metadati associati alla risorsa che garantiscono l'identità e integrità della risorsa veicolata da uno o più file)
- ✓ **Fruibilità** (vi sono sufficienti metadati associati alla risorsa che garantiscono l'accesso da parte della "comunità di riferimento")

Architettura MD - alta stabilità nel tempo

- ✓ File contenitore
 - ✓ WARC -- dati
 - ✓ XMLTape --- metadati
- ✓ File di controllo
 - ✓ Registro (log degli "eventi")
 - ✓ Anagrafe (chi sono gli "agenti" -- software e persone che operano per MD)

Architettura MD - alta adattabilità nel tempo

- ✓ Memorizzazione (storage):
 - ✓ usare gli oggetti più semplici e più diffusi sul mercato
 - ✓ ridondanza, multisito ecc.
- ✓ Indici di accesso e di controllo:
 - ✓ sempre ricostruibili da WARC e XMLTAPE
 - ✓ possono usare le tecnologie più efficienti disponibili

Sito BNCF

Nodo 1 -
Primario

Elemento 1.1

Elemento 1.2

Nodo n -
Replica

Elemento nR.1

Elemento nR.2

Sito BNCR

Nodo 1 -
Replica

Elemento 1R.1

Elemento 1R.2

Nodo n -
Primario

Elemento n.1

Elemento n.2

Dark Archive

DP: cosa si può fare oggi