



WP9 – Digitalizzazione del patrimonio storico



UOS di Rende



Agenzia per l'Italia Digitale
Presidenza del Consiglio dei Ministri

Workshop S&TDL - ROMA 5/11/2014



Consiglio Nazionale delle Ricerche

WP9: Ci stiamo occupando di ...

1. Digitalizzazione del patrimonio di interesse storico-scientifico
2. Indicizzazione dei contenuti per la costruzioni di vocabolari di dominio:
 - T2K – estrazione automatica di Named Entities (persone, organizzazioni, ecc.) per la creazione di:
 - Authority lists
 - Soggetti/parole chiave
3. Descrizione archivistico-documentale: XDams



Digitalizzazione del patrimonio di interesse storico-scientifico

- ✓ Risorse documentali:
 - Centro Nazionale Universitario per il Calcolo Elettronico (CNUCE)
 - Commissione Generale per l'Informatica (CGI)

- ✓ I documenti digitalizzati sono circa 300:
 - Provenienti dall'archivio centrale del CNR
 - Provenienti da archivi personali di dipendenti dell' IIT-CNR di Pisa

- ✓ La dimensione di ciascun file varia da un minimo di 50 KB a un massimo di 141 MB

- ✓ Risoluzione dei documenti 600x600 dpi
 - Qualità della digitalizzazione compromessa dalla scarsa qualità del documento originario



Digitalizzazione del patrimonio di interesse storico-scientifico

- ✓ Documenti CNUCE – CGI ancora da digitalizzare:
 - Materiale fotografico (circa 350 foto, che ritraggono personale CNR al lavoro e calcolatori IBM, datate 1971-1975)
 - Documenti presenti presso l'archivio generale dell'Università di Pisa
 - Video (DVD: SIRIO Prelaunch show)



Indicizzazione dei contenuti per la costruzione di vocabolari di dominio

- ✓ Sono state attuate le seguenti fasi:
 1. Conversione di un campione significativo di documenti (100 docs) del fondo CNUCE/CGI da .pdf in formato .txt
 - OCR (ABBYY)
 2. Estrazione automatica di termini rilevanti dal corpus creato
 - Uso del tool Text2Knowledge - T2K (ILC-CNR)
 3. Revisione dei termini estratti
 - Alta percentuale di termini generici restituiti da T2K
 - Arricchimento manuale di termini specifici per il dominio
 4. Mappatura a Nuovo Soggettario e a liste di dominio esistenti
 - Es. Authority list Persone mappata a VIAF



Indicizzazione dei contenuti per la costruzione di vocabolari di dominio

✓ Output della fase 2 (T2K):

- Glossario di termini rilevanti usato come base per la creazione di una lista di soggetti e parole chiave
 - Totale di 1109 termini (generici e di dominio)
- Lista di Named Entities usata come base per la creazione di Authority lists separate per:
 - Persone (~ 500 voci, in fase di espansione)
 - Lista arricchita con dati quali: varianti del nome, affiliazione, ruolo, dati anagrafici, mapping al record in VIAF, ecc.
 - Organizzazioni (250 voci)
 - Da integrare con liste standardizzate (es. ISTAT, MEF, MIUR, ANVUR, VIAF)
 - Luoghi (solo 65 voci)
 - Accordi con Regesta per uso in Xdams di Geonames o LinkedGeoData



Descrizione archivistico-documentale

- ✓ Descrizione dei documenti attraverso XDAMS v. 2.0 (v. giugno 2014)
 - Piattaforma di gestione documentale XML che sfrutta la piattaforma Extraway® XML Engine
 - Piattaforma per la conservazione, organizzazione, condivisione e valorizzazione dei patrimoni archivistici
 - Modello dati basato su elementi dello standard EAD
 - Conforme agli standard ISAD, ISAAR, FIAF (video) e Scheda F (foto)
 - Attributi dei metadati definiti dallo standard Dublin Core
- ✓ Funzione di “dialogo” da xDams verso la piattaforma DL in fase di definizione

Descrizione archivistico-documentale

The screenshot displays the xDams web interface. At the top, there is a navigation bar with the xDams logo and a search bar. Below this, a breadcrumb trail indicates the current location: "Benvenuto taverniti - Maria Taverniti - (livello: 1) sei in: Archivi / Archivio Storico". A toolbar shows "visualizza 100 elementi per pagina" and a refresh icon.

The main content area is divided into two columns. The left column shows a tree view of the document hierarchy. The right column displays the metadata for the selected document.

Document Hierarchy (Left Column):

- [Fondo Centro Nazionale Universitario per il Calcolo Elettronico] 05 luglio 1965
 - [Verbale n. 2 del 6 aprile 1965] Info Documento
 - Elementi inferiori collegati: 299
 - Posizione all'interno del ramo: 4
 - Livello di profondità: 1
 - Elemento collegato con data minore: "Calcolatore elettronico IBM 7090 - Parte generale" 1964 - 1965
 - Elemento collegato con data maggiore: [Fondo Centro Nazionale Universitario per il Calcolo Elettronico] 05 luglio 1965
 - [Atti generali]
 - [Atti repertoriati]
 - [Materiale a stampa]
 - Comunicazione da parte della Commissione Generale per l'Informatica
- [Commissione Generale per l'Informatica] 1982
 - [Verbali della Commissione] 1982
 - "Scioglimento Commissione" 37312" 29 febbraio 1980
 - "Elenco delle carte che si trasmettono alla Direzione Centrale Affari scientifici e programmazione". Protocollo n. 5463 23 dicembre 1980
 - "Commissione Generale per l'Informatica (C.G.I.). Normativa tariffaria per l'erogazione di servizio di calcolo". Protocollo n. 13470 06 febbraio 1980

Metadata (Right Column):

- SELEZIONA SPOSTA RIORDINA ELIMINA TAGLIA COPIA STAMPA**
- CODICE INTERNO E LIVELLO**
001.001.002 fondo
- DENOMINAZIONE E ESTREMI CRONOLOGICI**
[Fondo Centro Nazionale Universitario per il Calcolo Elettronico] , 05 luglio 1965
- SOGGETTO PRODUTTORE**
INFO
Consiglio nazionale delle ricerche (CNR)
- NOTE**

Navigation Bar (Bottom):

- SCHEDA STRUTTURA MODIFICA MULTIPLA TEST
- SUCCESSIVO SUPERIORE INFERIORE
- INSERISCI MODIFICA MODIFICA XML XML

Pianificazione delle attività future...

- ✓ Finalizzazione della customizzazione della piattaforma xDams
 - Funzione di importazione automatica di liste di autorità
 - prevedere un lookup special verso geonames o simili
 - Funzione creazione nuovo archivio
 - Profilazione utente
 - Personalizzazione dei lookup rispetto alla lista degli autori disponibili per quel profilo utente.
 - Creazione di un archivio soggetti/temi rispetto al soggettoario di Firenze, con creazione delle voci automatiche e poi associarle.
- ✓ Arricchimento delle liste di autorità e della lista di soggetti
- ✓ Conversione delle liste di autorità in LOD (Linked Open Data) per facilitarne il riuso e l'integrazione con altre risorse



S&T SCIENCE AND TECHNOLOGY DIGITAL LIBRARY



Maria Taverniti
maria.taverniti@cnr.it



Agenzia per l'Italia Digitale
Presidenza del Consiglio dei Ministri

ROMA 5/11/2014



Consiglio Nazionale delle Ricerche