Regione Emilia-Romagna

Calcolo automatico dell'indice di interoperabilità dei formati

Regione Emilia-Romagna

ParER - Polo archivistico Emilia-Romagna

parer
POLO ARCHIVISTICO
DELL'EMILIA-ROMAGNA

CC BY NC SA

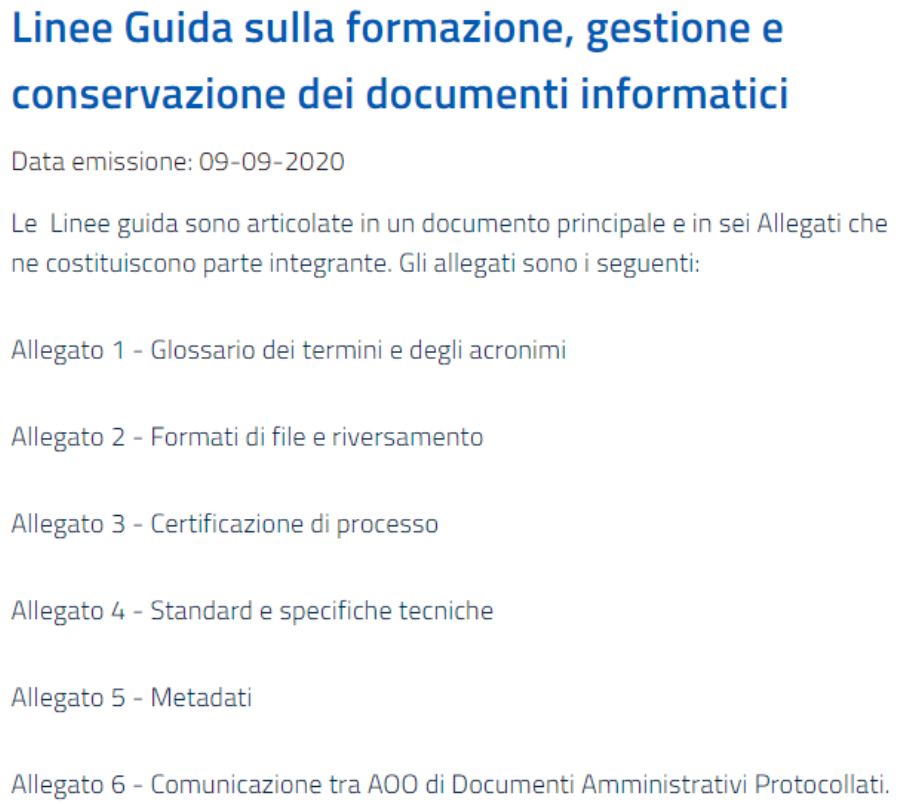
Introduzione

- Il Polo Archivistico Regione Emilia Romagna (ParER) garantisce che i file versati siano preservati all'interno dei propri archivi in modo sicuro, accessibile e conforme alle normative nel corso del tempo.
- È fondamentale garantire che i documenti conservati siano leggibili anche al passare di diversi anni. Per farlo:
  - È necessario riconoscere e gestire il formato di un documento al *momento del versamento*
  - È essenziale *una valutazione periodica* dei formati dei file conservati rilevando quelli a rischio di obsolescenza tecnologica.

Il problema dei formati

Il problema dei formati – Allegato 2 LLGG Agid

- L'Allegato 2 delle LLGG AGID «Formati di file e riversamento»
- Il documento è di 154 pagine e contiene la descrizione di 124 formati, divisi in 16 categorie
- par. 3.1 le PPA possono usare formati diversi da quelli elencati effettuando una valutazione di interoperabilità
- 3.2 modello semplificato quantitativo, ad ogni formato di file viene associato un valore numerico chiamato indice di interoperabilità (0-20, livello minimo 12).



Linee Guida sulla formazione, gestione e conservazione dei documenti informatici

Data emissione: 09-09-2020

Le Linee guida sono articolate in un documento principale e in sei Allegati che ne costituiscono parte integrante. Gli allegati sono i seguenti:

Allegato 1 - Glossario dei termini e degli acronimi

Allegato 2 - Formati di file e riversamento

Allegato 3 - Certificazione di processo

Allegato 4 - Standard e specifiche tecniche

Allegato 5 - Metadati

Allegato 6 - Comunicazione tra AOO di Documenti Amministrativi Protocollati.

Il problema dei formati – quanti sono?

- DELOS, File Formats Typology and Registries for Digital Preservation (2004)
  Progetto sponsorizzato dalla Commissione Europea e realizzato presso l'Università degli Studi di Urbino.
- Il documento di 54 pagine del 2004 categorizza i formati e suggerisce criteri di valutazione dei formati per l'uso da parte degli archivi.
- I formati individuati sono più di 15.000

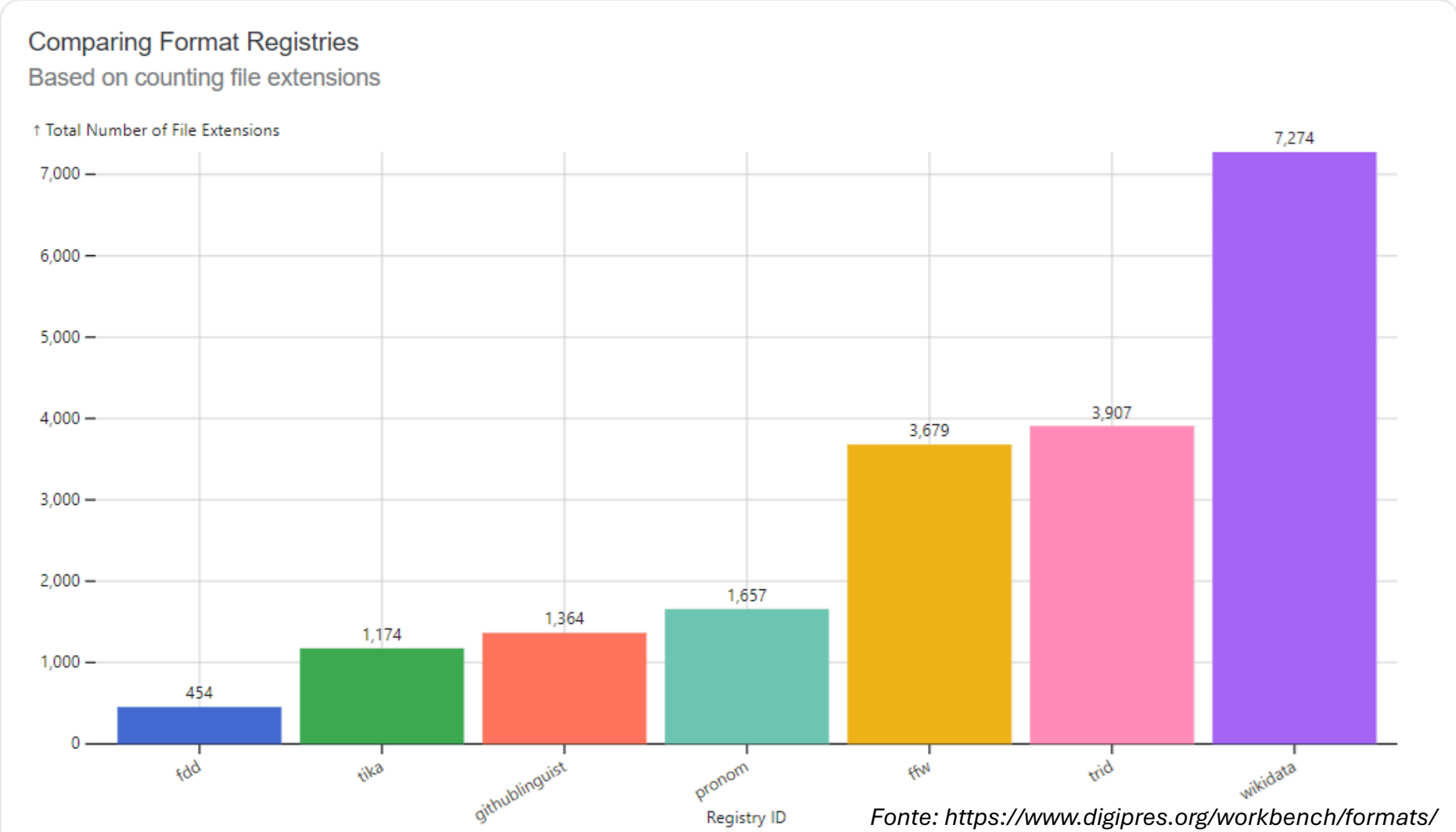
Il problema dei formati – esistono dei registri?

- La community DigiPres ha lanciato nel 2024 il progetto delle buone pratiche
- Per i formati sono stati indicizzati, analizzati e confrontati dati provenienti da diverse fonti di informazioni sui formati (noti anche come registri di formati) e strumenti di identificazione dei formati
- Ma per usarli è necessario trovare un modo per combinarli.

Fonte: https://www.digipres.org/workbench/formats/

Registro	Homepage	Descrizione
fdd	LoC FDD	Library of Congress Format Descriptions
ffw	FFW	'Just Solve The Problem' File Format Wiki by ArchiveTeam
pronom	PRONOM	PRONOM Format Registry by the National Archives (UK)
tika	Apache Tika	The Apache Tika Format Identification & Metadata Extraction Tool
trid	TRiD	The TRiD Format Identification Tool
githublinguist	GitHub Linguist	Linguist Format Identification Tool by GitHub
wikidata	WikiData	The WikiData Informatics Project

Il problema dei formati – una comparazione



Fonte: https://www.digipres.org/workbench/formats/

Fonte: https://it.wikipedia.org/wiki/Portable_Document_Format

Il problema dei formati – come riconoscerli?

- L'identificazione dei formati può avvenire tramite
  - l'estensione es. «.pdf»
    - Il sistema operativo capisce dall'estensione quale programma utilizzare per aprire il file
  - il mime type es. «application/pdf»
    - Rilevabile da software di identificazione
  - il magic number es. «%pdf»
    - Particolare sequenza di byte posta in genere all'inizio del file
- In realtà questi sistemi non sono esaustivi per identificare la versione di un formato (e ci sono pure i dialetti!)

Estensione	`.pdf`
Magic number	`%PDF`
Tipo MIME	`application/pdf`
Uniform Type Identifier (UTI)	`com.adobe.pdf`
Type code (Mac OS)	`PDF`
Sviluppatore	Adobe
1ª pubblicazione	15 giugno 1993
Ultima versione	2.0 (dicembre 2020)
Tipo	Formato documentale
Esteso a	PDF/A, PDF/E, PDF/UA, PDF/VT, PDF/X, PDF/VCR
Standard	ISO 32000-2
Formato aperto?	Sì
Sito web	www.iso.org/standard/75839.html e www.adobe.com/products/acrobat/adobepdf.html

Il problema dei formati - L'IA può aiutare?

- È possibile riutilizzare come base di conoscenza le informazioni sui formati già raccolte da fonti autorevoli?
- Lo stato dell'arte delle tecniche di Intelligenza Artificiale e Machine Learning possono aiutarci nella gestione del problema della valutazione dei formati e del calcolo dell'indice di interoperabilità?

Lo studio

- Nel 2023 ParER ha avviato uno studio congiunto con la sezione AI & Data di Engineering Group.
- Lo studio si è concluso nel 2024 con la produzione del documento Studio per la valutazione di un indice di interoperabilità dei formati da cui sono tratte le successive slide.

Lo studio – Il sistema di conservazione SacER

- Esiste un registro delle tipologie di formato gestite che riporta i fattori generali di interoperabilità definite dall'allegato 2.
- ha una forma tabellare, contiene 187 tipologie di formati ed è utilizzato per una valutazione del formato che abilita la successiva fase di versamento.
- L'identificazione del formato del file al versamento è basilare: tramite l'analisi del mimetype viene fatta una verifica di coerenza con il formato dichiarato dal soggetto versante.
- Non c'è riconoscimento della versione del formato del file né di eventuali sue caratteristiche specifiche.
- Tra i formati conservati da ParER, il 38% ha estensione PDF, il 30% è un XML, il 10% EML. Le restanti 184 tipologie di formati censiti occupa il 22% (286 milioni su 1,3 miliardi di file)

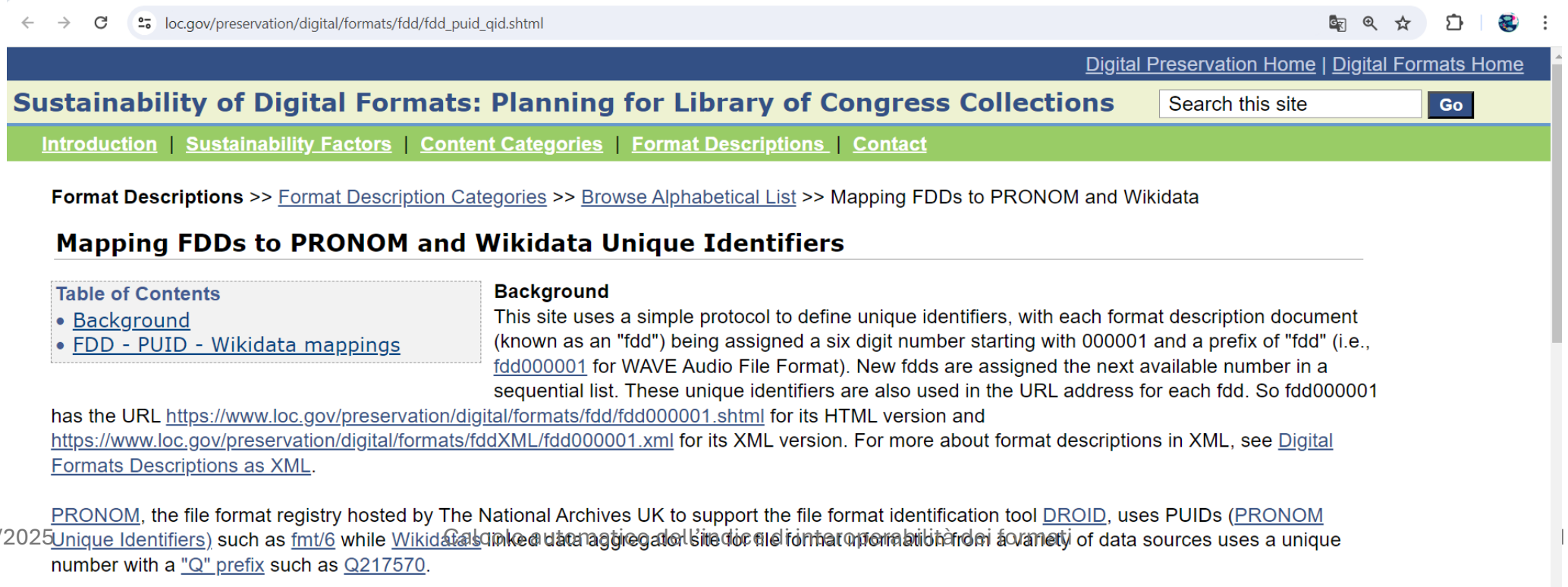
Lo studio – Allegato 2 LLGG AGID

- Introducono un modello per la valutazione dell'interoperabilità dei formati
  - suddividendoli in categorie generali e specifiche/ speciali che possono dipendere da fattori specifici del singolo file
- I criteri di interoperabilità hanno classi di punteggio a cui sono attribuiti dei valori numerici
  - Es. Per il criterio Natura, relativo alla standardizzazione del formato, le classi possono essere nessuno (0), de facto (2), de iure (3)

WORD® 2007	FORMATO DI FILE
Nome completo	WordProcessingML OOXML Extension
Estensione/i	.docx, .dotx
Specializzazione di	XML imbustato dentro ZIP
Tipo MIME	application/vnd.openxmlformats-officedocument.wordprocessingml.document application/vnd.openxmlformats-officedocument.wordprocessingml.template
Sviluppato da	Microsoft Corporation; ISO; ECMA
Tipologia di standard	proprietario (libero), estendibile, *de facto*, *testuale*
Livello metadati	3
Derivato da	Office Open XML; Microsoft® *Word*®
Revisione	11.1 (2018)
Riferimenti	▪ Microsoft, *Word extensions to ooxml (.docx) file format* v11.1 (2018) ▪ officeopenxml.com, *Anatomy of a WordProcessingML file*
Conservazione	Sì, solo profilo Strict; cfr. §2.8
Racc. per la lettura	Generico con riconoscimento obbligatorio
Racc. per la scrittura	Vedasi capoversi 10 e 11 per la conservazione.

Lo studio - Best practices internazionali

- Diversi organismi internazionali che si occupano di archiviazione e accesso alle risorse digitali hanno affrontato il problema
- tra i più rilevanti LocFDD, PRONOM, Wikidata



Lo studio – Loc FDD

- Loc è una fonte significativa per la creazione del catalogo dei formati e il calcolo automatico dell'indice di interoperabilità
  - Informazioni strutturate
  - accessibile in formato xml e AP
  - Interoperabile (Linked Open Data)
  - Manutenuto
- I criteri di sostenibilità della LoC raggruppano le caratteristiche dei formati per stimare la fattibilità e il costo della conservazione dei contenuti a fronte dei futuri cambiamenti tecnologici.
- i criteri di interoperabilità AGID hanno lo stesso obiettivo e richiedono requisiti simili.
- È possibile creare una mappatura tra i criteri di sostenibilità dei formati nel Catalogo LoC e i criteri generali di interoperabilità definiti da AGID.

Lo studio – Un catalogo dei formati

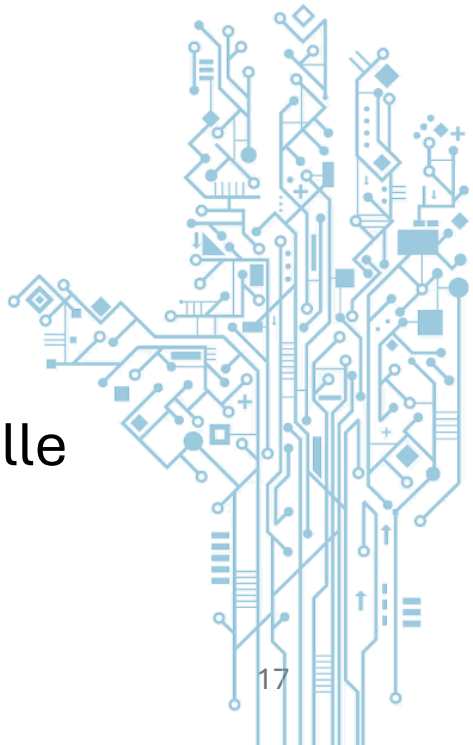
- Per riutilizzare le informazioni già raccolte da fonti autorevoli serve un'architettura a supporto del processo di valutazione dell'interoperabilità
  - Queste vanno raccolte, verificate, normalizzate e strutturate in modo da essere mantenute nel tempo.
- Il primo passo è creare un catalogo dei formati, un repository in cui sono memorizzate le informazioni necessarie al calcolo delle metriche utili alla valutazione dell'interoperabilità.

Lo studio – L’indice di interoperabilità

- Le informazioni presenti nell’All2 e nella LoC sono espresse esclusivamente in linguaggio naturale.
- Vanno elaborate e convertite in una rappresentazione numerica, in modo che un algoritmo possa calcolare l’indice di interoperabilità sui formati e sul singolo file.
- Le LLGG definiscono dei criteri generali che non dipendono dai fattori specifici del singolo file
  - la codifica dell’informazione può essere realizzata in una fase precedente alla fase di versamento del singolo file.

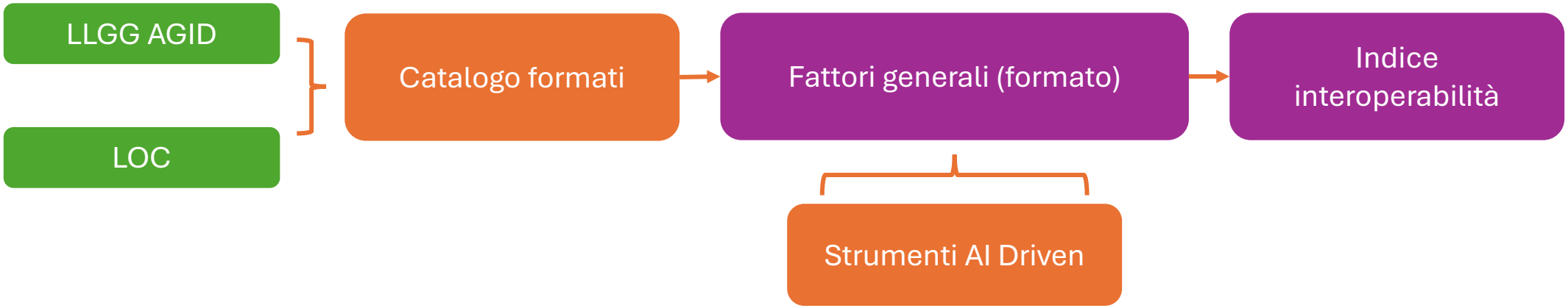
Lo studio – vantaggi dell’uso di IA

- valutazione delle proprietà del sistema tramite tecniche standardizzate, indipendenti dalla valutazione soggettiva dell’operatore umano
- riproducibilità dei criteri di calcolo punteggio
- possibilità di estendere il calcolo del punteggio a nuovi formati, in modo automatizzato, con una minima supervisione da parte dell’operatore
- possibilità di estendere la valutazione dei formati a banche dati diverse, che descrivano, tramite testo libero, le caratteristiche del formato stesso
- possibilità di una gestione multilingua omogenea del contesto e delle caratteristiche di formato



La soluzione tecnica

- Un primo intervento di realizzazione prevede la creazione:
  - di un prototipo semplificato di catalogo dei formati, in cui importare le informazioni prese da LoC
  - di un workflow per il calcolo automatico dell'indice di interoperabilità dei formati presenti in LoC utilizzando tecniche di alta efficienza basate sull'uso di AI



La soluzione tecnica – quale tecnologia?

- I criteri di valutazione forniti da Agid si dividono in classi di punteggio, ed ogni classe di punteggio è definita da un set di requisiti di valutazione.
- In questo scenario, il modello di AI ha il compito di ricondurre i criteri di sostenibilità LoC ad una delle classi di punteggio dei Criteri Agid.
- Per automatizzare questo processo, serve una strategia algoritmica che tenga in conto dei seguenti fattori:
  - Gestione del linguaggio naturale: sia i criteri Agid che le valutazioni LoC sono espressi in linguaggio naturale, un tipo di dato che presenta alta variabilità.
  - Comprensione e sintesi dei concetti: le informazioni contenute nei testi necessitano di essere individuate e sintetizzate ai fini della valutazione.
  - Comprensione del testo in lingue diverse: le classi di punteggio Agid sono espresse in lingua italiana, mentre le valutazioni LoC sono espresse in lingua inglese.

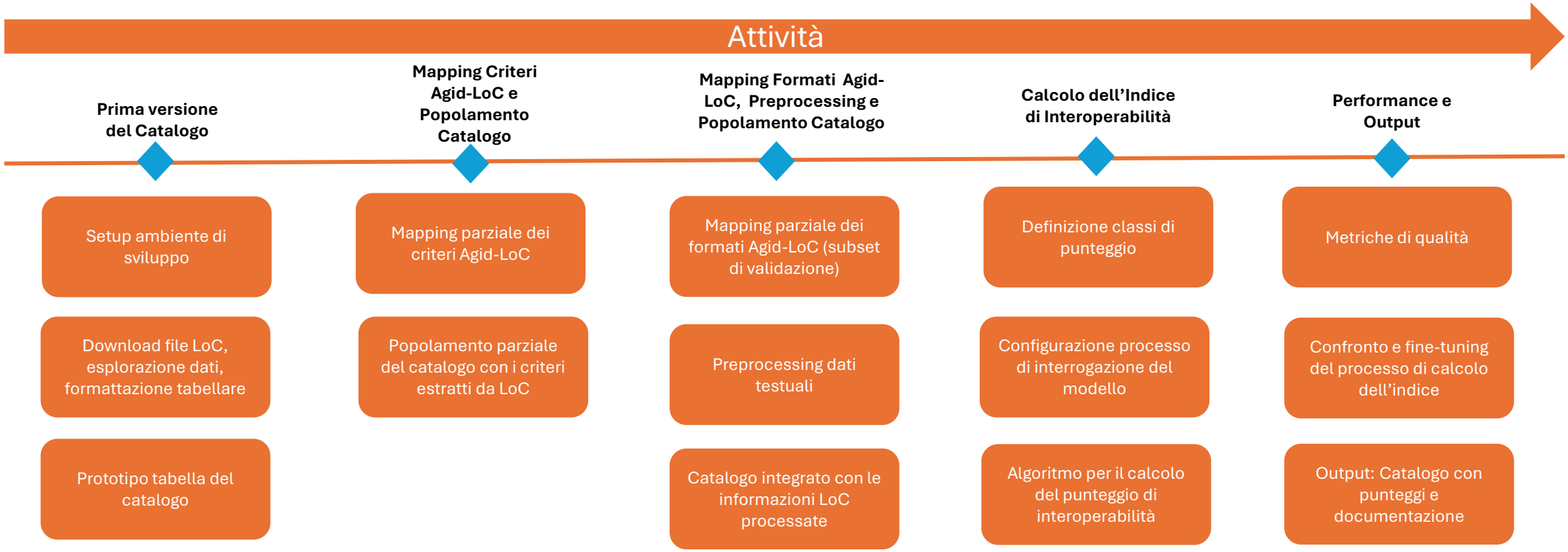
La soluzione tecnica – modello AI generativa

- Per ricondurre i criteri di sostenibilità LOC alle classi di punteggio AGID abbiamo bisogno di gestire linguaggio naturale, multilingua, comprendere e sintetizzare i concetti
- Il modello scelto è un modello di AI generativa di EngGPT, un framework AI di modelli linguistici Large Language Models (LLM) proprietario di Engineering per fornire strumenti di Generative AI in modo sicuro e affidabile:
  - È una soluzione customizzabile, può essere specializzato sul dominio di riferimento.
  - È fruibile come servizio *in cloud,* o disponibile *on premises*
  - Offre agli utenti un'opzione privata e italiana per acquistare i servizi GenAI e garantire il massimo controllo dei dati e della privacy
  - Può implementare policy di sicurezza e di *safety* che permettono un maggior controllo sulle risposte generate, diminuendo il rischio di risposte inesatte, offensive o pericolose.
  - Consente approcci di Explainable Artificial Intelligence, che aiutano a comprendere e rendere ancora più affidabili risultati ed output generati mediante algoritmi.

La realizzazione

- Novembre 2024 – Febbraio 2025
- Sviluppo con metodo Agile, approccio focalizzato sull'obiettivo di consegnare al cliente, in tempi brevi e frequentemente software funzionante e di qualità.
  - team di sviluppo piccolo, poli-funzionale e auto-organizzato
  - sviluppo iterativo e incrementale
  - pianificazione adattiva
  - coinvolgimento diretto e continuo del cliente nel processo di sviluppo.
- La prima fase di lavoro si è conclusa nel Febbraio 2025 con il rilascio del prototipo del catalogo dei formati in xls e del documento *Soluzione Tecnica per il Calcolo dell'Indice di Interoperabilità dei formati digitali Fase 1*» da cui sono tratte le slide a seguire.

La realizzazione - Il processo



Creazione dataset – estrazione LOC

- I formati estratti sono 571, ognuno ripetuto nel dataset con le informazioni relative ai criteri e valutazioni testuali.
- Per ogni formato sono state estratte le informazioni relative all'estensione, al mimetype, al magic number e le le informazioni relative ai fattori di sostenibilità LoC con le valutazioni testuali corrispondenti.
- Il dataset costruito sui dati della LoC è formato da 3976 righe

Formati LoC

Id	Name	Estensione	Mimetype
fdd000348	ISO Disk Image File Format	iso	application/x-iso9660-image
fdd000348	ISO Disk Image File Format	iso	application/x-iso9660-image
fdd000348	ISO Disk Image File Format	iso	application/x-iso9660-image
…	…	…	….

Criteri LoC

Criterio	Criterio Descrizione
disclosure	There is no comprehensive single….
adoption	Widely adopted. A variety of tool…
licensingAndPatents	None identified…
…	…

Creazione dataset – preparazione del dataset

- I criteri LOC sono descritti da un testo in linguaggio naturale in inglese. È stata necessaria una fase di preprocessing e pulizia.
  - Rimozione di caratteri di rumore o spazi extra
  - Rimozione di tag necessari solo per l'impaginazione del contenuto in un file XML
  - Analisi della distribuzione dei criteri estratti: La distribuzione dei criteri su tutti i file XML non è uniforme. Ad esempio, il criterio Technical protection mechanisms è presente in 564 formati su 571 totali.
  - Valutazioni ambigue come ad esempio la sigla «TBD» che potrebbe avere diversi significati;
  - Valutazioni mancanti come la stringa vuota o la stringa pari a «None»;
  - Valutazioni con rimandi ad altri formati che in genere iniziano con l'espressione «See <fddLink id=»;

Creazione dataset - LOC vs ParER

- Per logiche di semplificazione, è stato identificato un subset dei dati LoC al fine di ottenere un set di validazione per il confronto dei punteggi di Parer con quelli calcolati dal sistema.
- L'intersezione tra i formati ParER e LoC è stata fatta tramite la coppia «estensione» «mimetype»:
  - Mimetype
    - Attributo nel registro Parer: NM_MIMETYPE_FILE
    - Attributo nel dataset LoC: internetMediaType
  - Estensione
    - Attributo nel registro Parer: NM_FORMATO_FILE_STANDARD
    - Attributo nel dataset LoC: filenameExtension
- Il Subset di validazione comprende 175 record distribuiti su 25 formati.
- Obiettivo: analizzare quali sono i punti in comune e le differenze tra i due sistemi di valutazione attraverso delle metriche oggettive, in modo da identificare gli elementi oggetto di miglioramento in una successiva fase di fine tuning.

Classi punteggio -Mapping criteri AGID-LOC

- La mappatura realizzata è di tipo 1:1, ad un singolo criterio AGID corrisponde un singolo criterio LoC.
- È parziale, non tutti i criteri AGID trovano una corrispondenza coi criteri LoC.
  - Agid introduce il concetto di classe di punteggio. LoC, invece, identifica un set di criteri che danno indicazioni sul grado di preservazione di un contenuto digitale e la valutazione si basa sul grado di rispondenza rispetto ai desiderata espressi dal criterio ma non è misurabile.
  - I Criteri LoC e criteri Agid a volte valutano aspetti diversi dello stesso concetto.
    - Un esempio è la mappatura tra il criterio “Robustezza (lett. f)” e il criterio “Technical Protection Mechanisms”. Il criterio Agid si concentra perlopiù sul concetto di “grado di corruzione di un file e recupero delle informazioni non corrotte”, mentre il criterio LoC è più incentrato sull’aspetto di “grado di protezione dei contenuti del file” e sulla presenza o meno di sistemi di sicurezza.
  - 3 su 9 criteri AGID non sono stati valutati (ma solo il criterio estendibilità contribuisce al calcolo dell’indice di interoperabilità)

Criterio AGID	Descrizione Criterio AGID	Punti	Criterio LoC	Descrizione Criterio LoC
Natura (a)	Indica se il formato è considerato il formato standard da adottare per un certo tipo di contenuto, in base all'uso da parte degli utilizzatori, o in base a normative che ne regolano l'adozione in determinati contesti d'uso definiti	0,2,3	**Adoption**	Grado in cui il formato è già utilizzato dai principali creatori, distributori o utenti di risorse informative. Questo include l'uso come formato principale, per la distribuzione agli utenti finali e come mezzo di interscambio tra sistemi
Apertura (b)	Presenza o meno di specifiche tecniche, che illustrino i processi di creazione, di lettura e di utilizzo del formato (operational patterns)	0,3	**Disclosure**	Grado in cui esistono e sono accessibili le specifiche complete e gli strumenti per validare l'integrità tecnica a chi crea e mantiene contenuti digitali. La conservazione a lungo termine dei contenuti digitali non è possibile senza comprendere come l'informazione è rappresentata (codificata) come bit e byte nei file digitali.
Presenza di brevetti (c)	Specifiche del formato circa le sue caratteristiche e compatibilità con formati precedenti o altre tipologie, rilasciate dall'organizzazione privata o pubblica responsabile della sua proprietà intellettuale.	0,2,3,4	**Impact of Patents**	Presenza di brevetti e quindi termini di licenza
Estendibilità (d)	Possibilità di estendere le funzionalità del formato attualmente esistente	0,2	-	-
Livello modello metadati (e)	Livello del formato secondo il Modello per i Metadati definito dalle Linee Guida AGID	0,1,2,3	**Self-documentation**	La capacità di un formato digitale di contenere (in forma trasparente) metadati oltre quelli necessari per la semplice visualizzazione del contenuto nell'ambiente tecnico odierno
Robustezza (f)	Presenza o meno di meccanismi che permettano di valutare l'integrità del file	0,4	**Technical Protection Mechanism**	Misure appropriate per preservare i contenuti digitali e renderli accessibili alle future generazioni
Dipendenza dal dispositivo (g)	Richiesta di specifiche hardware e/o software per la creazione o lettura del formato	0,4	**External dependencies**	Le dipendenze esterne si riferiscono al grado in cui un particolare formato dipende da hardware, sistema operativo o software specifici per essere visualizzato o utilizzato, e alla complessità prevista per gestire queste dipendenze in futuri ambienti tecnici.
	-		**Transparency**	La trasparenza si riferisce al grado in cui la rappresentazione digitale è aperta all'analisi diretta con strumenti di base, inclusa la leggibilità umana usando un editor di testo.
Compatibilità (lett. h)	formati il cui standard prevede by design che un applicativo in grado di interpretare una data revisione possa anche leggere file formattati con revisioni precedenti (eventualmente entro un limite massimo) oppure in base a revisioni successive.	0/?		
Contenuto (lett. i)	Interpretazione del contenuto del file tramite esperto umano di dominio o algoritmo di parsing	0,0		

Classi Punteggio - la classe «non lo so»

- Per la gestione di casi in cui il modello potrebbe non riuscire ad assegnare alla valutazione testuale di un criterio LoC una classe di punteggio specifica è stata aggiunta la classe “non_lo_so” che ha punteggio sempre pari a 0
- Le casistiche, raccolte durante la fase di preprocessing, sono 3:
  - la valutazione testuale non contiene sufficienti informazioni o informazioni pertinenti al criterio in esame (es. *apparently not known to the compiler of this text, comments welcome*). Sottoclasse «no info»
  - la valutazione testuale è ambigua (riconoscibile da espressioni testuali come: *TBD, not known* e relative variazioni). Sottoclasse «non dedotto»
  - Il testo è mancante. Sottoclasse «no info»
- Comportamento conservativo: se il modello conversazionale non ha sufficienti elementi, potrebbe fornire informazioni non verificabili.

Configurazione modello

- Per la configurazione del modello di AI sono stati definiti due tipologie di template (insieme di informazioni che il modello può usare per la risoluzione del task della valutazione):
  - Human Template: parametri riferiti ad ogni richiesta di valutazione che viene inviata al modello
  - System Template: parametri riferiti alla descrizione del problema da risolvere, dell'output e dei comportamenti che ci si aspetta dal modello.

Human template	Nome Parametro	Descrizione
Richiesta	**Valutazione Testuale Criterio LoC**	Valutazione testuale del criterio LoC (in lingua inglese), estratta e preprocessata, a cui associare una classe dell'Allegato 2 Agid.
	Criterio Agid	Nome del criterio Agid, ottenuto dalla mappartura parziale, su cui eseguire l'assegnazione della classe.
Contesto	**Classi di Punteggio Agid**	Descrizione delle classi di punteggio relativamente al criterio da valutare.

System Template	Nome Parametro	Descrizione
Passi logici	**Definizione del task**	Descrizione del tipo di problema che il modello deve risolvere (in questo caso, la classificazione del testo LoC su una delle classi di punteggio disponibili).
	Specifica dell'output	Indicazioni sulla formattazione dell'output (es. il modello deve restituire un output come una stringa di testo o un output più strutturato, ad esempio sotto forma di dizionario). L'output deve contenere sia la valutazione che la motivazione.
	Gestione dei casi particolari	Indicazioni al modello sul comportamento da tenere per le valutazioni non utilizzabili (es. Unknown to the user).

Configurazione Modello - Contesto

- Semplificazione del linguaggio giuridico
- Arricchimento da diverse fonti
  - LLGG Agid
  - Linee guida nazionali per la valorizzazione del patrimonio informativo pubblico;
  - Definizioni dei fattori di sostenibilità della LoC

Criterio Natura (lett. a)

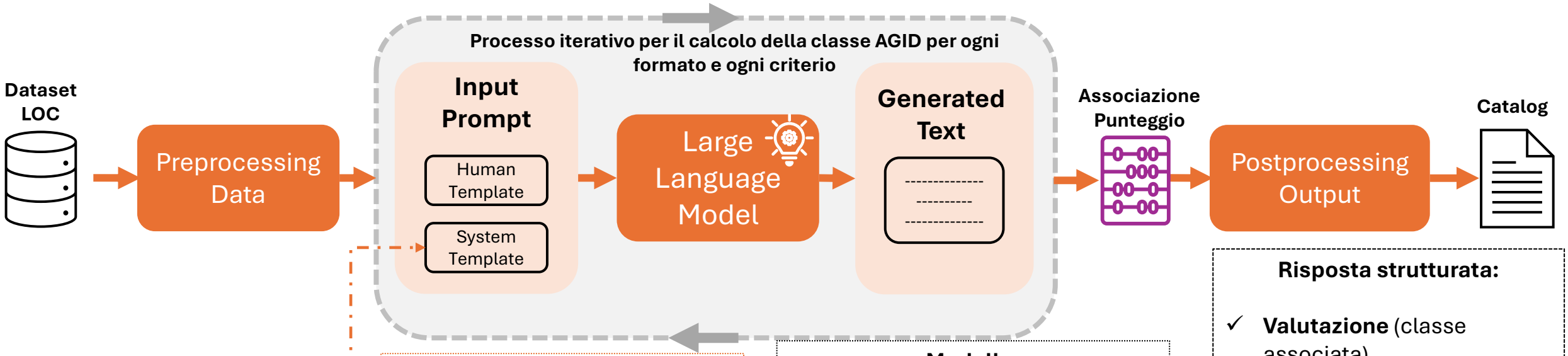
«Le classi per questo criterio sono le seguenti: de facto e de iure:

- *Un formato ha classe de facto quando questioni contingenti quali l'efficienza in casi d'uso reali, l'autoregolazione dei mercati di riferimento, l'efficacia tecnica, ne hanno determinano una larghissima e non trascurabile diffusione, per lo meno in settori di riferimento. L'adozione diffusa di un formato digitale può essere dimostrata dall'inclusione di strumenti nei pc, dal supporto nativo nei browser web o negli strumenti di creazione di contenuti leader di mercato, inclusi quelli destinati all'uso professionale, e dall'esistenza di molti prodotti concorrenti per la creazione, manipolazione o visualizzazione di oggetti digitali nel formato.*
- *Un formato può avere invece classe de iure, quando esistono normative che ne obblighino, o per lo meno ne raccomandino, l'uso in determinati contesti amministrativo-legali e settori di riferimento. Ad esempio, quei formati che esaminati da altre istituzioni archivistiche e accettati come formati archivistici prediletti.»*

In nero le definizioni dei criteri LLGG Agid.

In rosso le informazioni prese da LoC che vanno ad arricchire le definizioni «de facto» e «de iure»

Configurazione modello – processo di lavoro



Contesto:

- ✓ **Classi Agid** (definizioni prese da Allegato 2- Agid)
- ✓ **Criteri LoC** (informazioni rilevanti forniti dalle descrizioni dei criteri della LoC)

Modello:

- ✓ Il flusso algoritmico è stato sviluppato sulle **macchine Engineering Group** in **linguaggio Python**.
- ✓ Il modello usato per assegnare le classi Agid è un **modello di AI generativo basato su LLM** capace di comprendere e generare linguaggio naturale.

Risposta strutturata:

- ✓ **Valutazione** (classe associata)
- ✓ **Motivazione** (testo in italiano che spiega la scelta della classe, in italiano)
- ✓ **Sentence** (frase determinante che ha portato alla classificazione)
- ✓ **Punteggio** (punteggio associato in base a LLGG Agid)

Interrogazione modello

Vogliamo chiedere al modello di valutare il formato DOCX, Id LoC: fdd000397 criterio «Adoption» rispetto al Criterio Agid «Natura (lett.a)»

Interrogazione modello – Human template

Human template	Nome Parametro	Descrizione (rielaborazione)
Richiesta	**Valutazione Testuale Criterio LoC**	International open standard. Maintained by ISO/IEC JTC1 SC34/WG4. Originated by Microsoft Corporation and first standardized through ECMA International in 2006. Approval as ISO/IEC 29500 was in 2008
	Criterio Agid	Natura (lett.a)
Contesto	**Classi di Punteggio Agid**	**Le classi per questo criterio sono le seguenti:** • **de facto** quando questioni contingenti quali l'efficienza in casi d'uso reali, l'autoregolazione dei mercati di riferimento, l'efficacia tecnica, ne hanno determinano una larghissima e non trascurabile diffusione, per lo meno in settori di riferimento. L'adozione diffusa di un formato digitale può essere dimostrata dall'inclusione di strumenti nei pc, dal supporto nativo nei browser web o negli strumenti di creazione di contenuti leader di mercato, inclusi quelli destinati all'uso professionale, e dall'esistenza di molti prodotti concorrenti per la creazione, manipolazione o visualizzazione di oggetti digitali nel formato. • **de iure**, quando esistono normative che ne obblighino, o per lo meno ne raccomandino, l'uso in determinati contesti amministrativo-legali e settori di riferimento. Ad esempio, quei formati che esaminati da altre istituzioni archivistiche e accettati come formati archivistici prediletti.»

Interrogazione modello – System template

System Template	Nome Parametro	Descrizione
Passi logici	**Definizione del task**	Sei un agente esperto nella valutazione dei formati digitali. Il tuo compito è di valutare la {richiesta}, in lingua inglese, che ti viene passata in input dall'utente e di associarla ad una delle classi di punteggio riportate nel {contesto}
	Specifica dell'output	Come output devi restituire un json strutturato, così composto: {Valutazione: richiesta dell'utente passata al modello, Criterio: criterio Agid che va a valutare la richiesta dell'utente, Classe : classe associata, Motivazione : testo in italiano che spiega la scelta della classe, Sentence: frase determinante che ha portato alla classificazione}
	Gestione dei casi particolari	Se nel testo della valutazioni trovi espressioni simili ad "Unknown to the user", restituisci la classe «Non_lo_so»

Interrogazione modello - output

- Il testo della motivazione, generata dal modello di AI, mostra i due passaggi fatti dal modello per arrivare alla soluzione:
  - Il modello sintetizza le informazioni del testo LoC passato in input dalla richiesta
  - Basandosi sulle informazioni del Contesto, deduce che la disponibilità della documentazione è un elemento rilevante per l'associazione alla classe

```
{
    «Valutazione» : «International open standard. Maintained by ISO/IEC
                    JTC1 SC34/WG4. Originated by Microsoft
                    Corporation and first standardized through ECMA
                    International in 2006. Approval as ISO/IEC 29500
                    was in 2008»,

    «Criterio» : «Natura (lett.a)»,

    «Classe» : «de iure»,

    «Motivazione» : «Il formato DOCX ha formato de Iure per lo standard
                    ISO/IEC»,

    «Sentence» : «ISO/IEC 29500-1, Information technology -- Document
                    description and processing languages -- Office
                    Open XML File Formats --»

}
```

Il campo Sentence riporta la frase più significativa per il modello, ovvero quella che contiene le informazioni principali per la classificazione.

Un formato ha classe *de iure* , quando esistono normative che ne obblighino, o per lo meno ne raccomandino, l'uso in determinati contesti amministrativo-legali e settori di riferimento

Il prototipo del catalogo

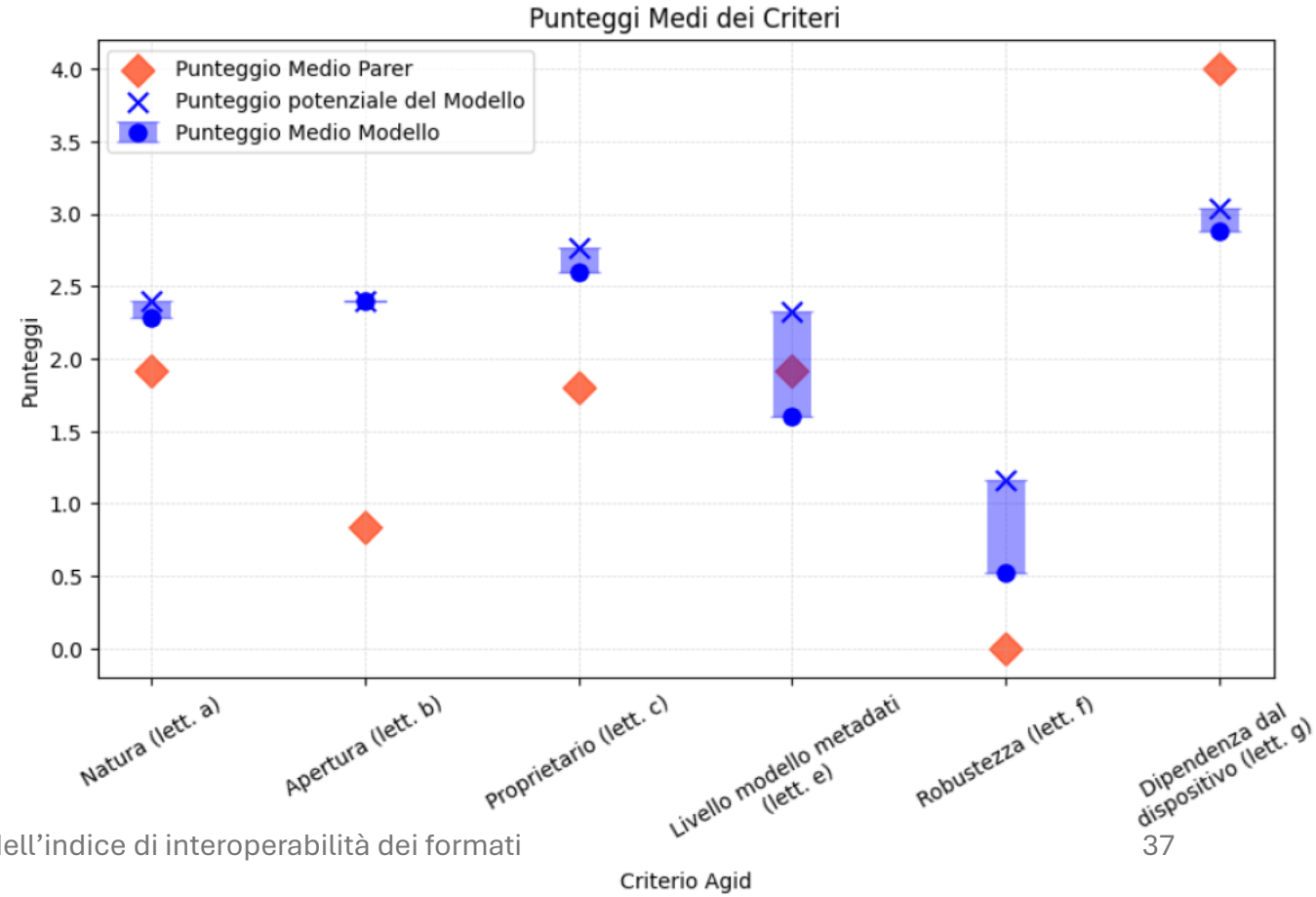
Id LoC	Nome Format	Titolo Formato	Estensione	Mimetype	Standard Fla	Standard	Standard Link	Standard Descript	Criterio Descrizione Original	Criterio Descrizio	Criterio	Criterio Agid	Classe	Sottoclasse non_lo_s	Motivazione	Sentence	Punteggio	Indice Interope rabil
fdd000397	DOCX/OOXML_2012	DOCX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5	docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document	1	ISO/IEC|ECM	https://en.wi	ISO/IEC 295(	Very widely used.	Very widely us	adoption	Natura (lett. a)	**de iure**		il formato ha classe de iure per lo standard ISO/IEC , ISO , ECMA	ISO/IEC 29500-1, Information technology - Document description and processing languages -- Office Open XML File Formats -- Part 1: Fundamentals and Markup Language Reference and ISO/IEC 29500-4, Information	3	15
fdd000397	DOCX/OOXML_2012	DOCX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5	docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document	0				International ope	International c	disclosure	Apertura (lett. b)	**aperto**		Il formato è un standard aperto, mantenuto da ISO/IEC JTC1 SC34/WG4 e originato da Microsoft Corporation. È stato	International open standard. Maintained by ISO/IEC JTC1 SC34/WG4. Originated by Microsoft Corporation and first	3	15
fdd000397	DOCX/OOXML_2012	DOCX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5	docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document	0				None beyond XML	None beyond X	externalDependencies							15
fdd000397	DOCX/OOXML_2012	DOCX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5	docx	application/vnd.openxmlformats-officedocument.wordprocessingml.document	0				The specification	The specificati	licensingAn atents							15

WORD® 2007 — FORMATO DI FILE

Nome completo	WordProcessingML OOXML Extension
Estensione/i	.docx, .dotx
Specializzazione di	XML imbustato dentro ZIP
Tipo MIME	application/vnd.openxmlformats-officedocument.wordprocessingml.document application/vnd.openxmlformats-officedocument.wordprocessingml.template
Sviluppato da	Microsoft Corporation; ISO; ECMA
Tipologia di standard	proprietario (libero), estendibile, *de facto*, *testuale*
Livello metadati	3
Derivato da	Office Open XML; Microsoft® *Word*®
Revisione	11.1 (2018)
Riferimenti	▪ Microsoft, *Word extensions to OOXML (.docx) file format* v11.1 (2018) ▪ officeopenxml.com, *Anatomy of a WordProcessingML file*
Conservazione	Sì, solo profilo Strict; cfr. §2.8
Racc. per la lettura	Generico con riconoscimento obbligatorio
Racc. per la scrittura	Vedasi capoversi 10 e 11 per la conservazione.

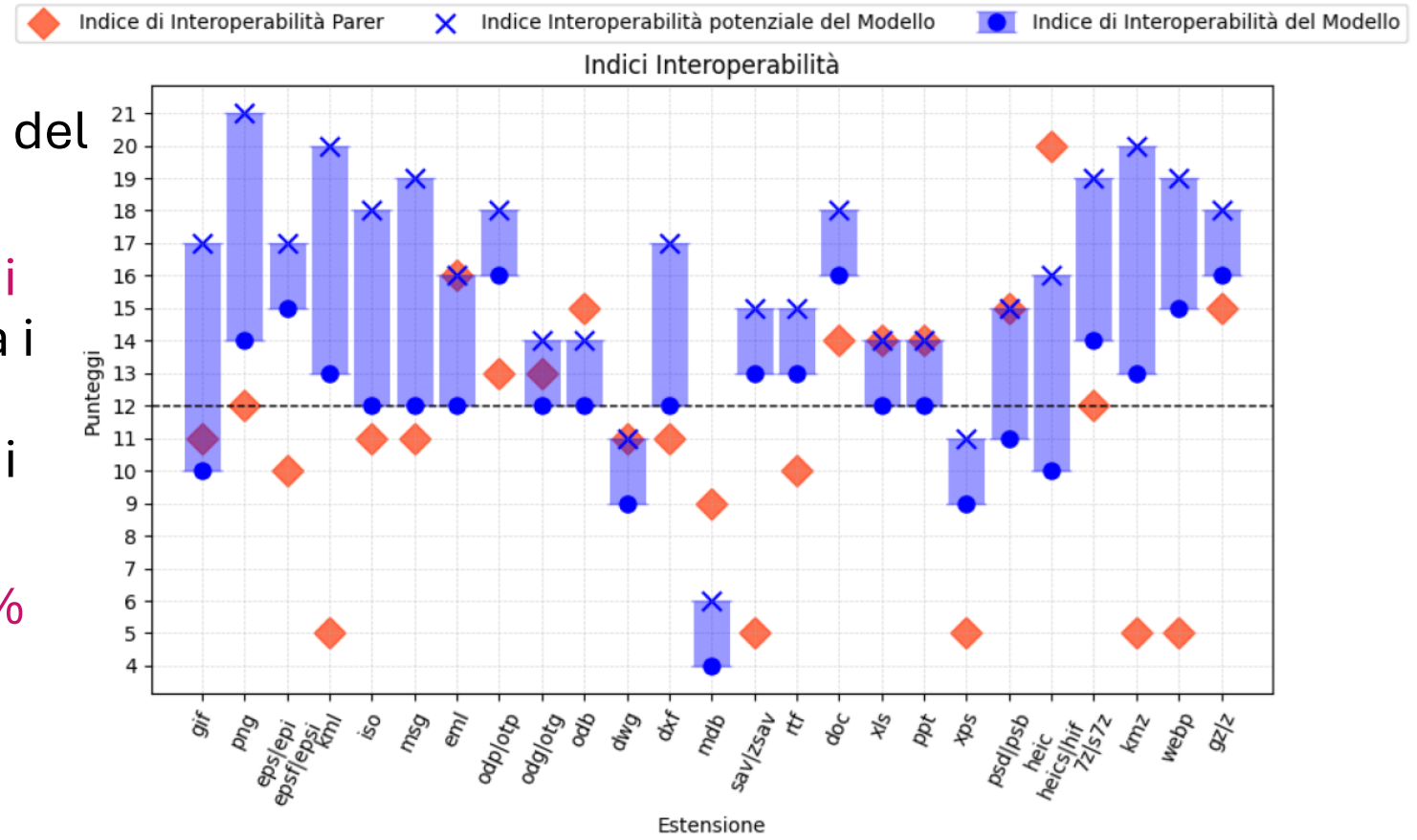
Metriche – analisi dei criteri agid

- Sono stati confrontati i punteggi medi per ogni criterio, calcolati sui 25 formati del subset di validazione.
- È stata definita la metrica “Similarità %”, che esprime la differenza tra i valori di punteggio medio del criterio rapportato al massimo valore possibile per quel criterio.
- In generale, la «Similarità %» assume un punteggio elevato, mediamente superiore al 65%. Questo indica che, in media, anche quando i valori non sono coincidenti, i punteggi forniti dal modello sono comunque vicini a quelli del registro PARER



Metriche – analisi indice interoperabilità

- Sono stati confrontati gli indici di interoperabilità calcolati sui 25 formati del subset di validazione.
- È stata definita la metrica “Similarità di Interoperabilità %”, come differenza tra i valori degli indici di interoperabilità in rapporto al massimo valore possibile di interoperabilità pari a 20.
- Lo score calcolato spesso supera l’80%



Metriche – esito delle analisi

- L'analisi dei punteggi ha evidenziato che il modello fornisce risultati con una buona Similarità % rispetto al registro PARER sia per i singoli criteri che per l'Indice di Interoperabilità.
- I casi di "non_lo_so" possano influenzare sul risultato finale dato che hanno sempre punteggio pari a 0.
- Il criterio estendibilità (0,2) non è stato valutato e ha sempre valore 0.
- Si potrebbero migliorare le valutazioni del modello integrando più informazioni nel parametro di Contesto fornito al modello di AI

Fine tuning

- La modalità di sviluppo AGILE ha consentito di effettuare alcune ottimizzazioni durante il ciclo di sviluppo in particolare sulla rielaborazione del contesto e la gestione della classe «non lo so».
- Una volta realizzata la prima versione del catalogo è stato strutturato un percorso di validazione e ottimizzazione organizzato in 11 task.

	Dati e fogli da consultare	Descrizione
1	Catalogo.xlsx	File Excel consegnato alla fine della Fase 1 con tutti i formati della LoC valutati secondo mappatura parziale con i criteri Agid.
2	SUE 32485 - export_registro_Formati.xls	Registro Parer fornito in Fase 1. Risulta utile confrontare i risultati del Catalogo con quelli del Registro Parer in modo da recuperare eventuali informazioni o linee guida consolidate da Parer.
3	Foglio "1 -Roadmap Fine Tuning"	Foglio del presente documento contenente attività suggerite con relative istruzioni da seguire al fine di verificare quali siano le informazioni da integrare nel sistema per implementare funzionalità di fine tuning.
4	Foglio "2 -Contesto"	Foglio del presente documento contenente la descrizione dei criteri secondo Agid ma rielaborate in modo che possano essere lette dal modello di AI per la valutazione dei criteri LoC. In questo caso è utile da consultare come suggerito per alcuni task del foglio "1 - Roadmap Fine Tuning".
5	Foglio "3 - Casi non_lo_so"	Foglio del presente documento contenente un subset del Catalogo di casi con classe "non_lo_so", in particolare casi in cui il modello non è riuscito a dedurre la classe (nella documentazione tecnica si fa riferimento alla sottoclasse "NON DEDOTTO")

Fine tuning – regole per il criterio Natura

- Le informazioni sullo Standard «de iure» sono state estratte dalla sezione «Documentation» delle schede formato LoC al fine di affindare l'associazione della classe «de iure» per il criterio «Natura (lett. a)».

Sustainability factors

Disclosure	Open standard.
Documentation	PNG (Portable Network Graphics) Specification, Version 1.2 (at http://www.libpng.org/pub/png/spec/1.2/PNG-Contents.html) ISO/IEC 15948:2004 Information technology -- Computer graphics and image processing -- Portable Network Graphics (PNG): Functional specification. W3C Portable Network Graphics (PNG) Specification (Second Edition), same text as ISO/IEC 15948:2004, at https://www.w3.org/TR/PNG/
Adoption	As of 2005, PNG was supported by most browsers. However, Internet Explorer 6.x for Windows did not support the transparency feature. Slow

Configurazione **parametrizzabile** che contiene la **seguente lista di standard**:

- ISO
- ISO/IEC
- OASIS
- W3C
- ECMA
- RFC

Estrazione dell'informazione sullo standard

Catalogo LoC

Catalogo arricchito con i seguenti campi:

- **Standard Description**: testo estratto dalla sezione «Documentation»
- **Standard:** il valore dello standard estratto secondo configurazione dal testo estratto.
- **Standard Flag**: valore 0 oppure 1. Se ha valore 1 in corrispondenza di «Natura (lett. a)» è stato trovato uno standard come da configurazione.

Fine tuning - Standard

- La lista di 6 standard implementata nel modello [ISO,ISO/IEC, OASIS, W3C, ECMA, RFC] è stata integrata in base a quanto riportato nei campi "Standard Description" e "Standard" della LOC»
- La lista di Standard de iure aggiornata contiene 15 standard da estrarre da LoC.
- Il Catalogo aggiornato presenta uno Standard de iure (campo «Standard Flag» = 1) per 192 formati LoC
  - 31 in più rispetto agli standard trovati in Fase 1.

	Standard	# Formati (su un totale di 571)
1	ISO	122
2	ISO/IEC	87
3	W3C	28
4	SMPTE	21
5	IETF	20
6	RFC	17
7	ECMA	14
8	OASIS	11
9	3GPP	4
10	ETSI	2
11	MIL-STD	2
12	ECH	1
13	MIL-SPEC	0
14	FIPS	0
15	FGCD	0

Fine tuning – regole per il criterio Apertura

- Distinzione classi "aperto" e "chiuso": verificare se un formato che ha per Natura la classe «de iure» si può considerare un formato con classe «aperto»
- La regola applicata, a valle della valutazione con il modello di AI, corregge i casi di formati valutati con classe «chiuso» per «Apertura (lett. b)» in «aperto» qualora per il criterio «Natura (lett. a)» risulti «de iure».
- In tal caso anche la Motivazione viene corretta con la seguente: «Il formato è aperto dato che è uno standard de iure.»

Id LoC	Nome Formato	Motivazione old	Classe Old	Classe New
fdd000388	EML	Il formato EML non ha una documentazione specifica disponibile, quindi viene classificato come chiuso.	chiuso	aperto
fdd000522	GeoPDF_file	Non esiste una documentazione dettagliata sulle funzioni del TerraGo GeoPDF Toolbar e sulle capacità del software TerraGo che utilizza il formato PDF.	chiuso	aperto
fdd000625	Apple_Fork	Il formato è proprietario e non è pubblicamente disponibile la documentazione specifica tecnica. Tuttavia, è disponibile alcune documentazione.	chiuso	aperto

Fine tuning - Valutazioni in corso

- Rivalutazione dei casi «non dedotto» con l'analisi delle informazioni presenti in altri campi LOC.
- Analisi del criterio «Transparency».
- Analisi dei dati relativi alle relazioni tra formati
  - 330 formati su 571 hanno il campo mime type nullo, di questi 175 hanno relazione del tipo «subtype of» con altri formati.
  - I formati con mimetype non nullo, e che sono in relazione di tipo «Subtype of» con altri formati, condividono lo stesso mimetype?
  - In caso positivo è possibile dire che i formati con mimetype nullo hanno lo stesso mimetype del formato gerarchicamente superiore?

Conclusioni

- La necessità di una standardizzazione dei criteri di idoneità alla conservazione riconducibile alle best practice internazionali
- Prossimi passi:
  - Arricchire il contesto del modello con altre fonti
  - Integrare un sistema di valutazione del formato del file al versamento nel sistema di conservazione
  - Integrare il modello con la valutazione dei fattori specifici del file versato

Grazie per l’attenzione

marianna.tascone@regione.emilia-romagna.it

Tutti i marchi, loghi e segni distintivi registrati e non registrati (e/o riferimenti di qualunque tipo) inseriti nel presente documento appartengono ai legittimi proprietari e sono pubblicati in osservanza delle normative vigenti.