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

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

viking lander data format file pr

google.com/search?q=viking+lander+data+format+file+problem&sca_esv=03f51f66e1b437d3&rlz=1C1GCEB_enIT1133IT1133&sxsrf=AE3TifOS0rfJnYj5dKhSlawIJ8LbX2x3Eg%3A17493126...

Google

viking lander data format file problem

Tutti Video Immagini Video brevi Notizie Web Libri Altro Strumenti

AI Overview

The primary "file format" problem with the Viking Lander data was not simply about the format itself, but about the degradation of the magnetic tape storage media and lack of sufficient metadata to understand the data for later generations. This led to the data becoming unusable as the tapes "brittled" and the original format became difficult to encode. Additionally, errors arose during data conversion from BCD to ASCII, causing loss of information.

Here's a more detailed breakdown:

1. Magnetic Tape Degradation:

- The Viking Lander data was stored on magnetic tape, which over time became "brittle" and cracked, making it difficult to read and potentially corrupting the data.
- This physical degradation was a major concern, as it meant the data could be lost forever if not preserved.

2. Lack of Metadata and Interpretability:

- The original data formats and their relationships to the scientific instruments and mission objectives were not fully documented in a way that was easily understood for future use.
- Some conversions of the data to other formats, like ASCII, also introduced errors, for example, by not properly handling binary headers in the original files.
- This lack of comprehensive metadata made it difficult for scientists to understand and interpret the data without extensive effort.

3. Data Conversion Issues:

- The data was initially stored in Binary Coded Decimal (BCD) format and later converted to ASCII by the National Space Science Data Center (NSSDC).
- During the conversion, errors occurred, particularly in how binary headers were handled, leading to a loss of important information contained within those headers.

Long Term Digital Preservation | Byrd CLS - HAI | Legal

Data from NASA's Viking missions to Mars in the 1970s was nearly lost to history. It was stored ...

HAI | Legal

University of Texas Institute for Geophysics Technical Report ...

This data set was originally generated from SEISF files, described below, and was written ...

Institute for Geophysics

Viking Lander Data | - Atlas of Digital Damages

The tapes went “brittle”, that is one reason why the data were not usable. But there is also mentioning of the impossibility to...

Atlas of Digital Damages

Mostra tutti

QUESTO ARTICOLO HA PIÙ DI 1 ANNO | FLASHES | Mercoledì 17 aprile 2024

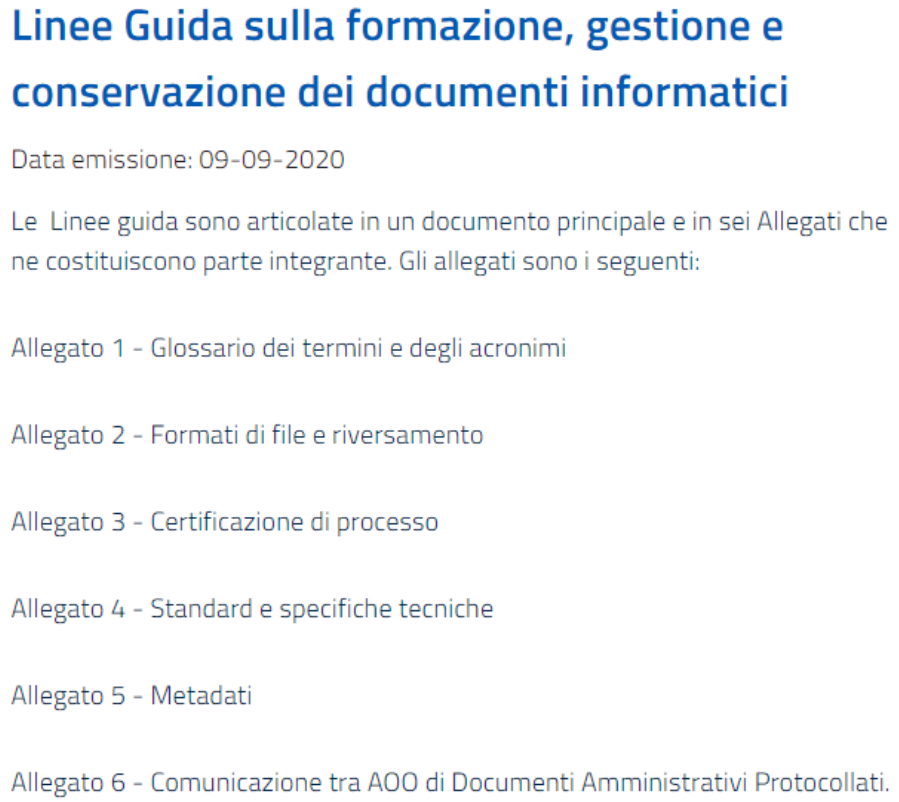
I documenti del ponte sullo Stretto inviati al ministero e definiti illeggibili sono letteralmente illeggibili

Tabella 5-15 Matrice OD urbana per Messina (veicoli nel giorno medio annuo) (scenario di progetto 2047)

]rqd#	Sruw#Wlhp hvwhu#Jdgd#Vdq#Iudqfhvfr# h#Wlhp hvwhu#		Srqwh#		Wrwdoh#	
	Ohjjhul#	Shvdqwl#	Ohjjhul#	Shvdqwl#	Ohjjhul#	Shvdqwl#
P ld#	3#	3#	::#	4<4#	::#	4<4#
Wlhp hvwhu#	3#	3#	47;#	576#	47;#	576#
Frqwhvvh#	3#	3#	4<:#	5;#	4<:#	5;#
Jd}}l#	3#	3#	47:#	46#	47:#	46#
Fdlrc#	3#	3#	4<3#	43#	4<3#	43#
Fhqwur#Vwrulfr#	3#	3#	543#	6<#	543#	6<#
Erffhwwd#	3#	3#	83#	3#	83#	3#
Jlrvwud#	3#	3#	98#	3#	98#	3#
Ulbjr#	3#	3#	45<#	4<#	45<#	4<#

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 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.

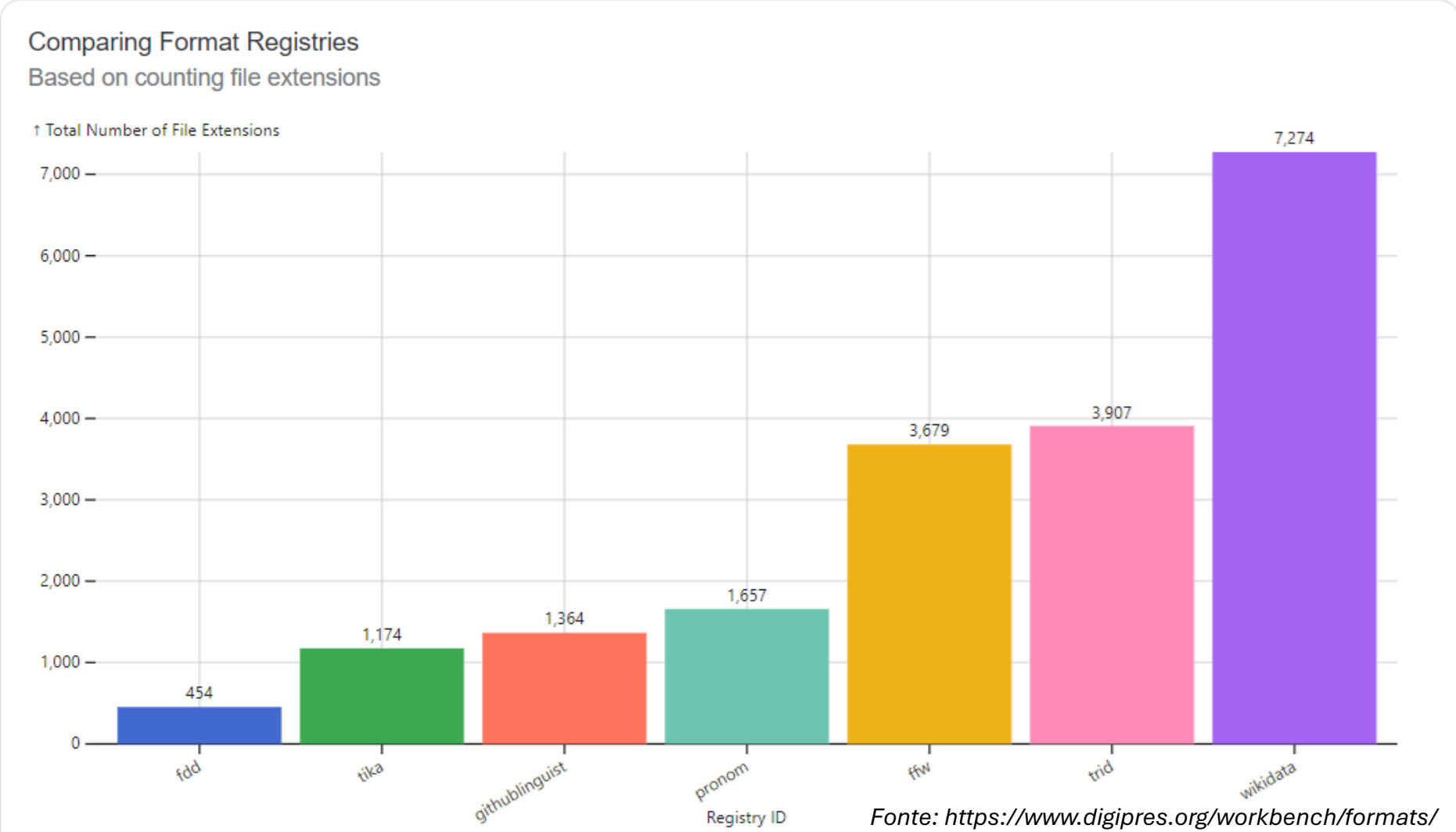
Il problema dei formati – esistono dei registri?

- 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?

Il progetto

- 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
- A Febbraio 2025 è stato rilasciato il prototipo del catalogo dei formati in xls e del documento Soluzione Tecnica per il Calcolo dell'Indice di Interoperabilità dei formati digitali Fase 1

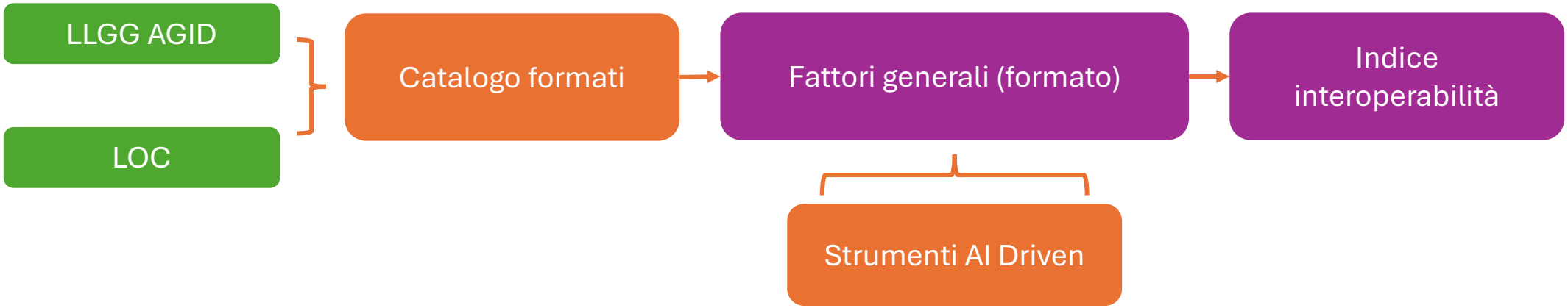
Lo studio

- Sistema di conservazione SacER
  - Registro dei formati su cui è stata fatta una valutazione di interoperabilità definite dall'all. 2
- Allegato 2 LLGG Agid
  - modello per la valutazione dell'interoperabilità dei formati in cui i criteri di interoperabilità hanno classi di punteggio di valore numerico
- Best Practices internazionali
  - Nella LOC sono presenti i criteri di sostenibilità in cui sono descritte le caratteristiche dei formati utili a stimare la fattibilità e il costo della conservazione dei contenuti

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.

La soluzione

- Il primo intervento di realizzazione ha prodotto:
  - un prototipo semplificato di catalogo dei formati, in cui importare le informazioni prese da LoC
  - un processo di calcolo automatico dell'indice di interoperabilità dei formati presenti in LoC utilizzando tecniche di alta efficienza basate sull'uso di AI



La realizzione

- ✓ Creazione Dataset
  - Estrazione di 571 formati da LOC per criterio di sostenibilità, 3976 record totali
  - pulizia e ottimizzazione
- ✓ Creazione subset di validazione
  - sottoinsieme di formati che potevano essere ricondotti ai formati presenti nel registro Parer, 175 record distribuiti su 25 formati.
- ✓ Mapping LOC-AGID
  - Mappatura di tipo 1:1 ma parziale, in alcuni casi sono presenti sfumature di significato diverse
  - 3 su 9 criteri AGID non sono stati mappati

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		

Interrogazione modello

Vogliamo chiedere al modello di valutare il formato DOCX, Id LOC: fdd000397 criterio «Adoption» rispetto al Criterio Agid «Natura (lett.a) e assegnare la classe di punteggio.

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**	Adoption refers to the degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a master format, for delivery to end users, and as a means of interchange between systems. If a format is widely adopted, it is less likely to become obsolete rapidly, and tools for migration and emulation are more likely to emerge from industry without specific investment by archival institutions. [...]

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 (+ 2)** 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 (+3)** 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 «de iure» che vale 3 punti.

```
{
    «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

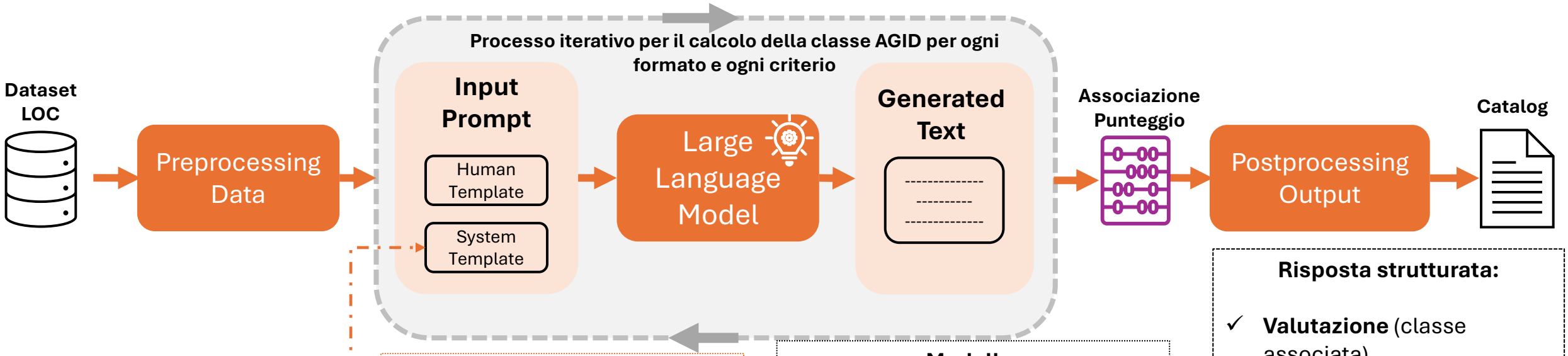
Il prototipo del catalogo

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.

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)

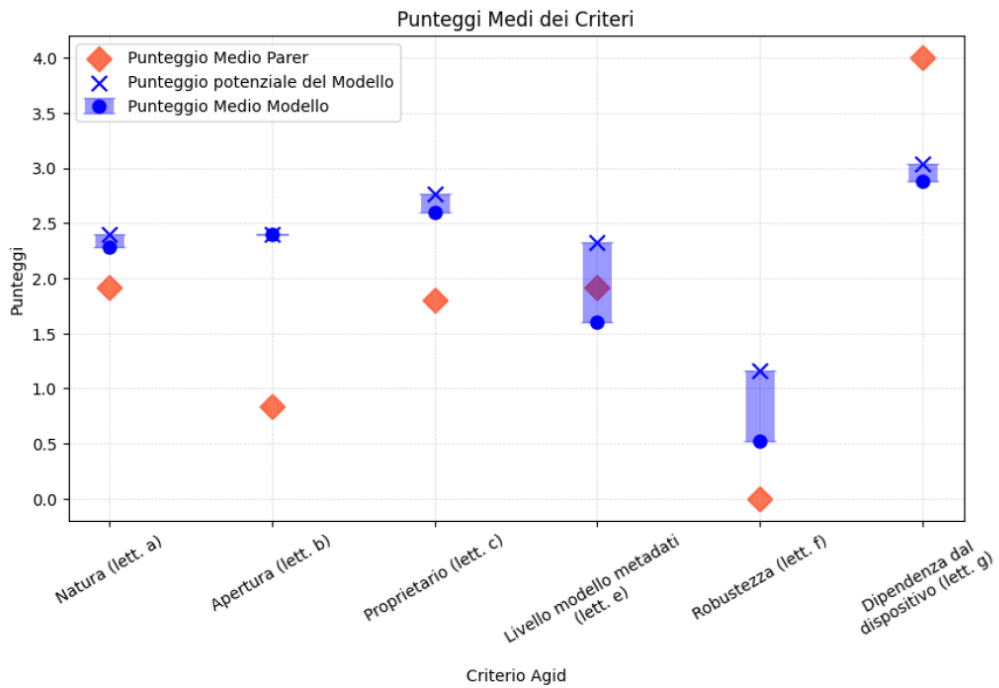
Modello:

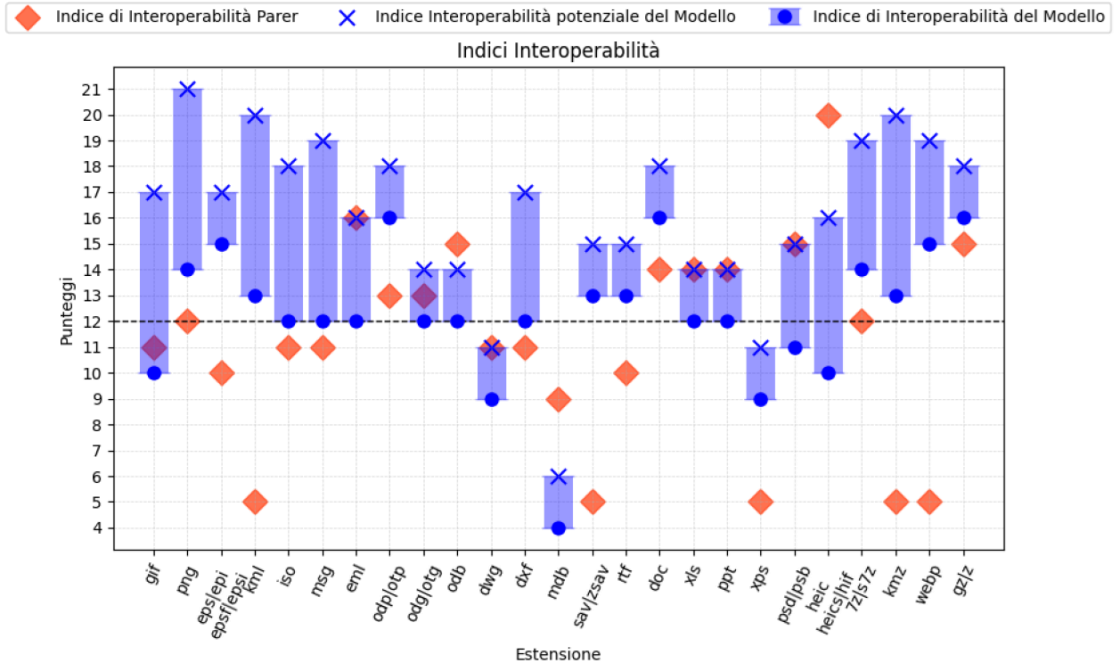
- ✓ Il flusso algoritmico è stato sviluppato sulle **macchine Engineering Group** in **linguaggio Python**.
- ✓ Il modello usato per assegnare le classi Agid è un **modello di AI generativo basato su LLM** capace di comprendere e generare linguaggio naturale.

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)

Metriche – analisi dei risultati





- Il modello fornisce risultati con una buona Similarità % rispetto al registro PARER sia per i singoli criteri che per l'Indice di Interoperabilità.
- 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")

Conclusioni

- La sfida dell'obsolescenza digitale è viva e concreta e il problema dei formati è solo uno degli aspetti da curare.
- La novità dell’indice di interoperabilità: un criterio di misurazione dell’idoneità alla conservazione
- Assenza di una standardizzazione internazionale dei criteri di conservabilità dei formati per una mappatura completa e affidabile.
- Prossimi passi: integrazione del catalogo in sacer e estensione del catalogo con registri come Pronom e Wikidata in moda da migliorare l'affidabilità del processo di conservazione.

Grazie per l’attenzione

marianna.tascone@regione.emilia-romagna.it