CONFERENZA GARR 2025
FRONTIERE DIGITALI

Layout Parser, come creare un dataset di qualità per allenare l'Intelligenza Artificiale

Silvano Imboden, Gabriele Marconi, Simona Caraceni, Rossella Pansini, Fauzia Albertin

Cineca

Il progetto MIC Digital Library

Infrastruttura

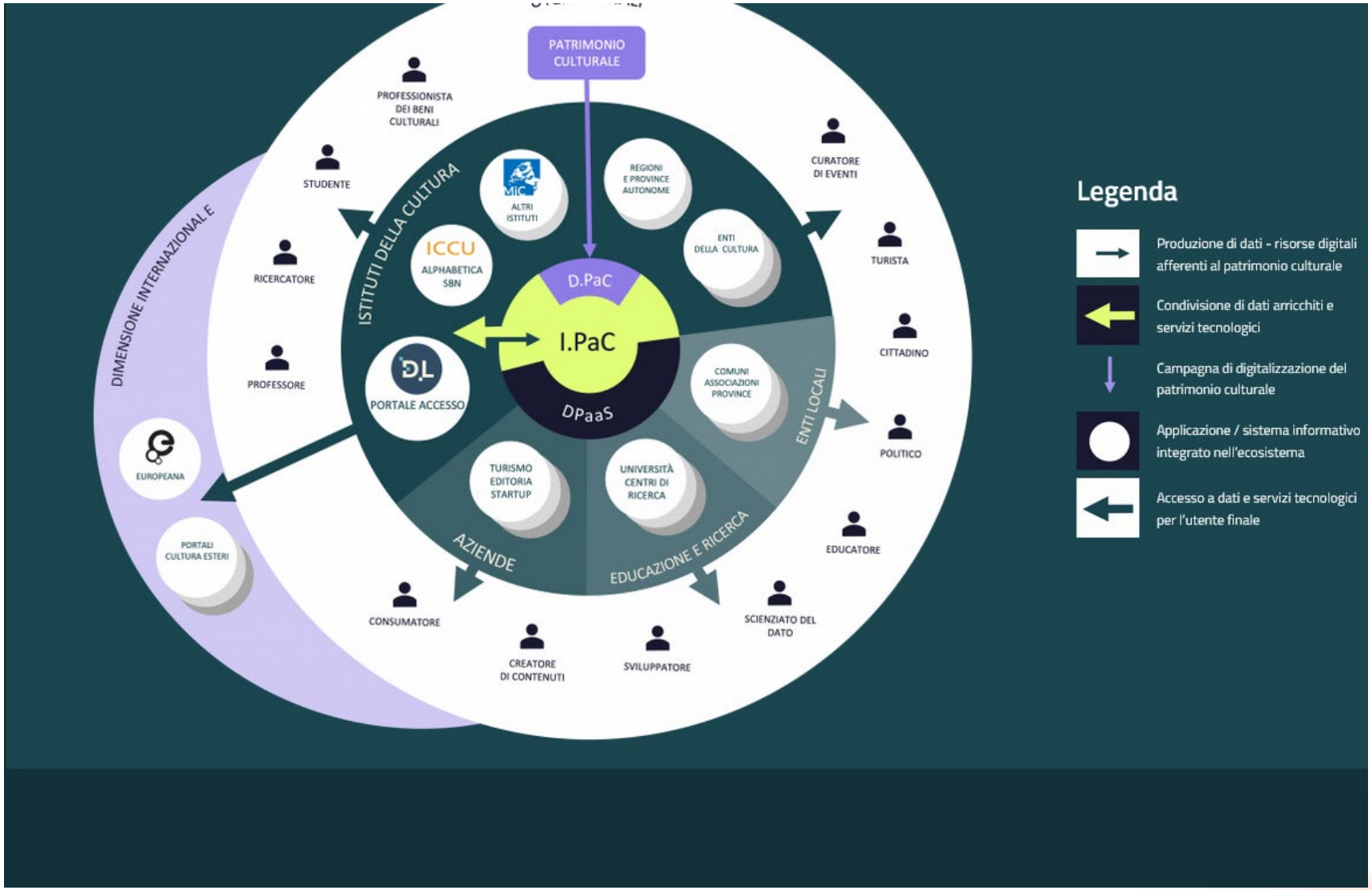
I.PaC, l'*Infrastruttura e servizi digitali per il Patrimonio Culturale*, è lo **spazio dei dati** progettato per conservare, gestire e arricchire il patrimonio culturale digitale del Paese, in linea con le principali strategie nazionali ed europee. Nasce dall'esigenza di superare la frammentarietà dei sistemi di fruizione e dal bisogno di gestire dati stratificati ed eterogenei per formato, tipologia, dominio di appartenenza e politiche di protezione, secondo modelli concettuali flessibili e in sicurezza.

I.PaC presenta un **complesso sistema di servizi digitali avanzati**, basati su tecnologie innovative orientate al *Cloud*. Implementa funzioni relative alla **gestione e all'arricchimento delle risorse digitali**, fondate sia su modelli e schemi predefiniti (motori a regole e ontologie) sia su algoritmi di intelligenza artificiale (AI), ed **espone un ampio catalogo di API di cooperazione applicativa** (in lettura e scrittura) relative a dati di dominio e cross-dominio.

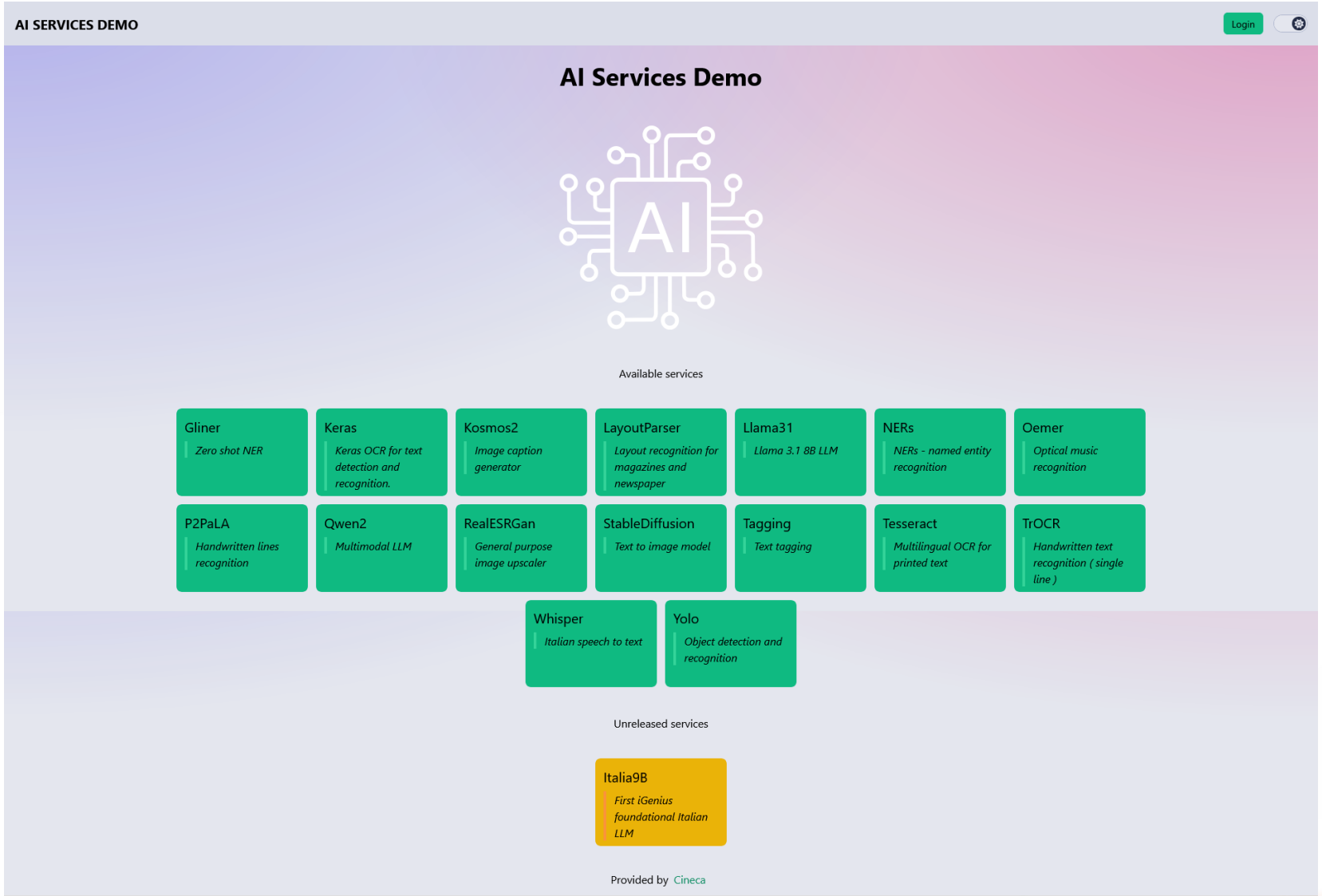
Scenario di riferimento

Lo sviluppo di I.PaC, e più in generale di un ecosistema digitale del patrimonio culturale italiano, rientra tra le azioni strategiche delineate nel *Piano nazionale di digitalizzazione del patrimonio culturale*, e si inserisce all'interno del più ampio progetto di trasformazione digitale promosso dalla Digital Library del Ministero della Cultura per il quinquennio 2022-2026.

Il progetto MIC Digital Library



La macchina delle cartucce di AI



https://aiws.hpc.cineca.it/

Layout Parser

https://aiws.hpc.cineca.it/

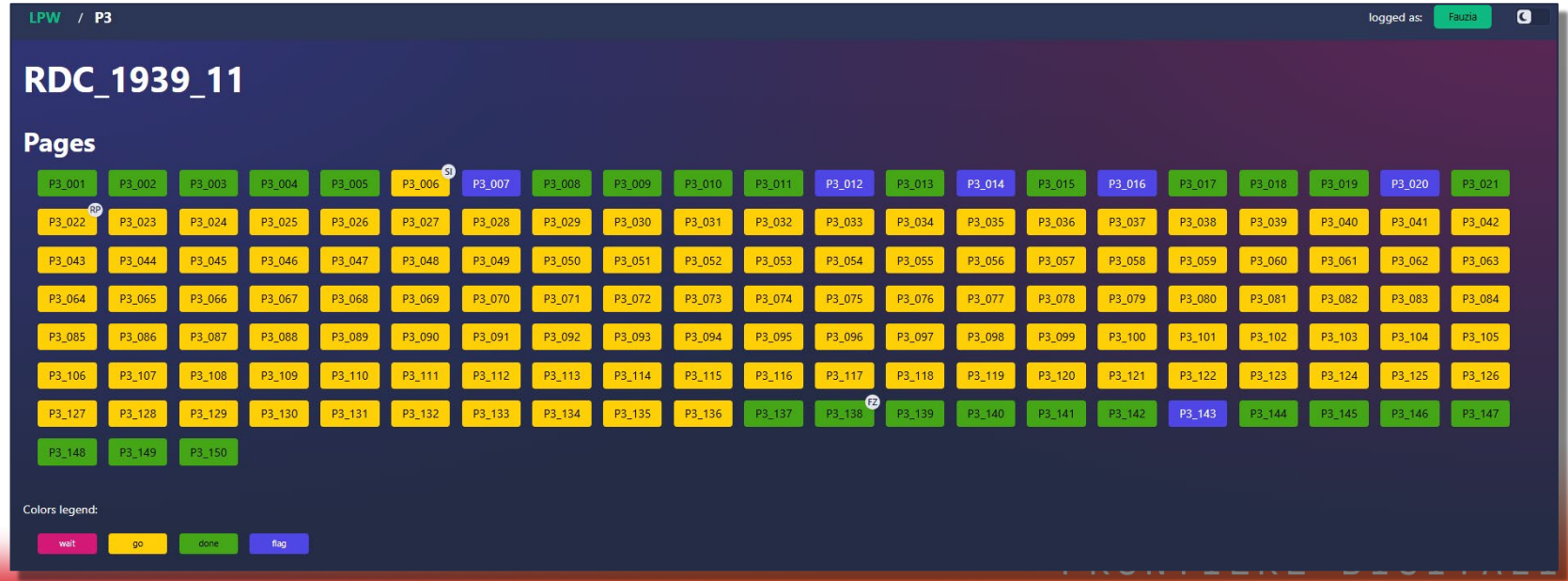
Istruzioni

Scopo della piattaforma è allenare il motore AI di Layout Parser ad identificare correttamente le differenti categorie di testo presenti all'interno di una pagina di quotidiano – **titolo, testo, fotografia, disegno** e **pubblicità**

Una volta entrati nella cartella di progetto ci troveremo davanti le varie pagine di quotidiano identificate da un codice e di colore differente

- Il codice identifica in modo univoco la pagine
- Il colore identifica le azioni che sono già state fatte su di essa:

wait	
go	La pagina può esser messa in lavorazione
done	Il lavoro è stato completato
flag	La pagina da attenzionare in quanto vi sono problematiche o dubbi



Istruzioni

Una volta entrati nella pagina che dobbiamo lavorare ci troveremo:

- l'immagine della pagina
- già evidenziate le aree identificate dal motore Layout Parser

Come si vede dall'immagine queste non sono corrette in quanto parte dei titoli e dei testi non è stato identificato in alcun modo

Scopo di questo lavoro è identificare TUTTE le aree della pagina secondo 5 cate

1	**Titolo**
2	**Testo**
3	**Fotografia**
4	**Disegno**
5	**Pubblicità**

Istruzioni

Nella parte superiore della pagina troviamo il nome della pagina che stiamo lavorando, il nostro user name e il pulsante delle impostazioni – in troveremo anche la corrispondenza colore-tipologia

La prima modalità di lavoro è **Edit Areas** – che ci permette di aggiungere, togliere o modificare le aree i

Istruzioni

Posizionando il mouse sull'area di interesse e contemporaneamente uno dei numeri da 1 a 5 il sistema identificherà automaticamente l'area assegnandogli la categoria

1 **Titolo**

2 **Testo**

3 **Fotografia**

4 **Disegno**

5 **Pubblicità**

La dimensione dell'area identificata automaticamente viene determinata dalle linee guida – evidenziate in giallo.

Nel caso in cui l'area identificata automaticamente non fosse corretta, questa va modificata:

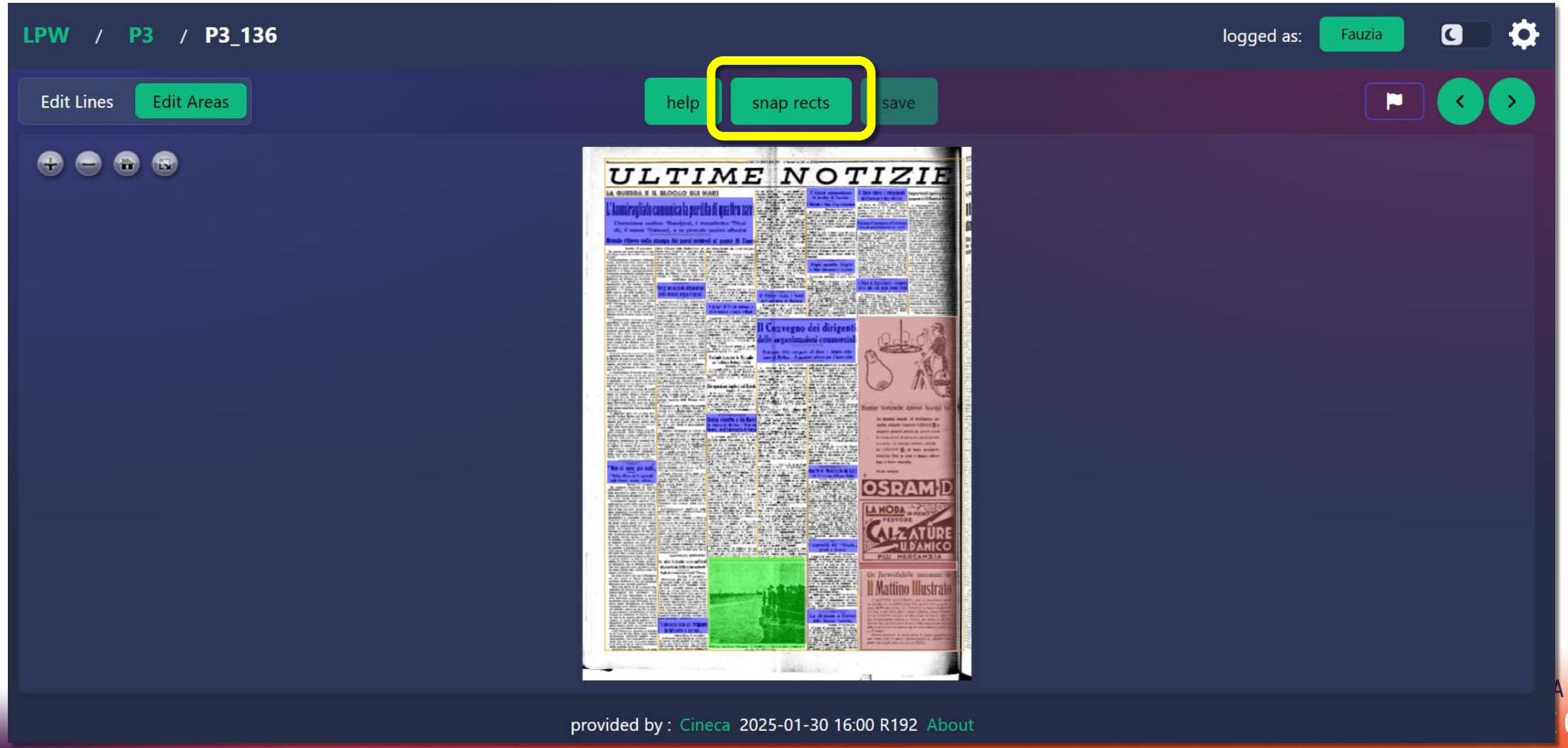
- agendo sui cursori ed allargandola o stringendola sino alla linea guida successiva.
- tenendo contemporaneamente premuto SHIFT è possibile muoversi al di fuori delle linee guida

E' possibile agire in ogni momento sulle aree identificate e compiere ulteriori azioni successive quali:

- **Split**: posso separare un area in due premendo S per lo split verticale o D per quello orizzontale. Apparirà una linea di divisione mobile che verrà confermata rilasciando il mouse
- **Merge:** selezionando due aree e premendo J le aree vengono unite in una unica

Istruzioni

Nel caso in cui si riscontrassero errori nelle linee guida PRIMA di cominciare il lavoro sulle aree è possibile chiedere al sistema di rettificare le linee agendo tramite il pulsante Snap rects

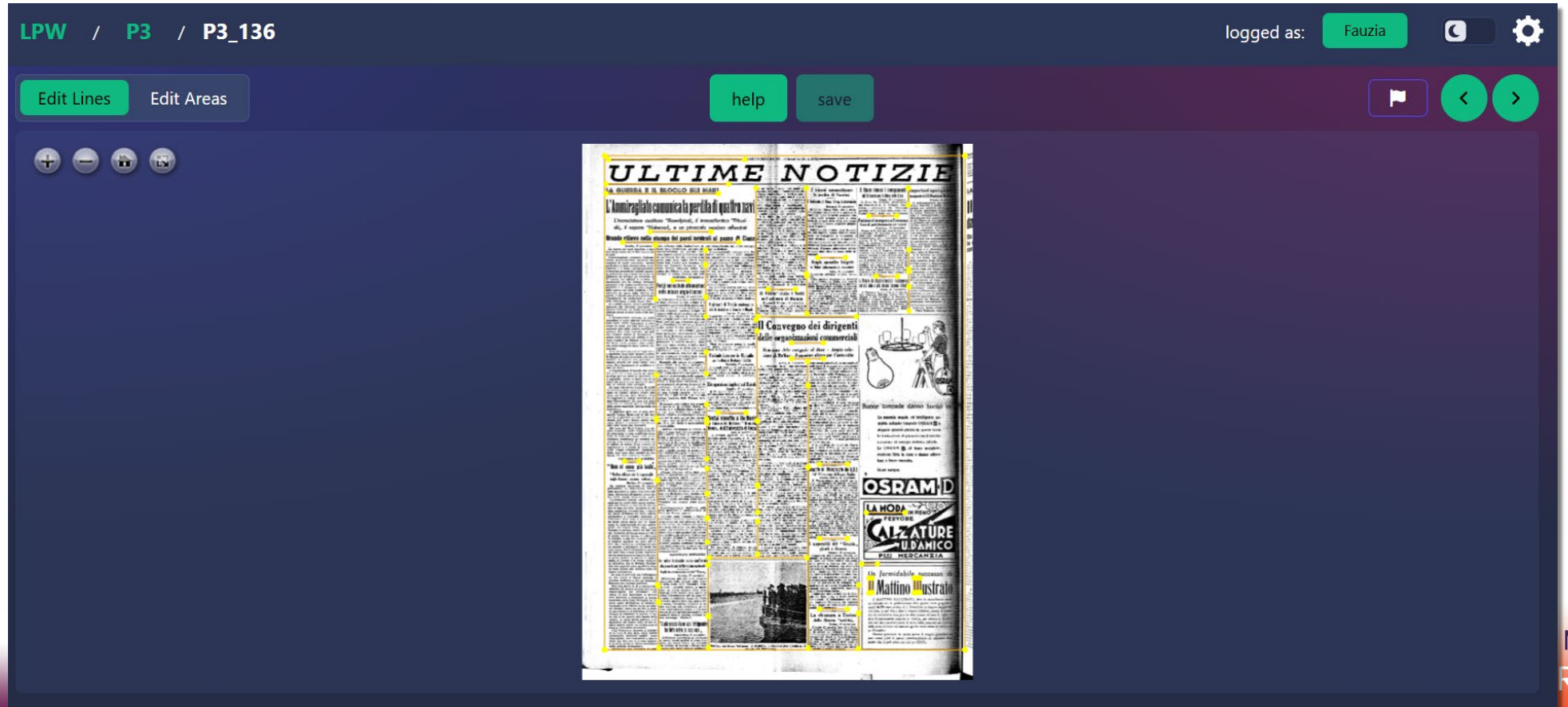


Istruzioni

Se gli errori sussistono è possibile entrare nella seconda modalità di lavoro: **Edit Lines**

In questa modalità è possibile:

- Spostare i marker di inizio e fine linea cliccando e trascinandoli
- Cancellarli cliccando e premendo CANC
- Creare nuovi marker cliccando e premendo CTRL



Alcune norme

Titolo


L'INCREMENTO DELLA PRODUZIONE

Le norme per gli orari di lavoro nel settore industriale

E' prevista una più larga e più intensa occupazione


I "Cento Anni„

di Giuseppe Rovani

- I titoli vanno suddivisi se provvisti di titolo e sottotitolo
- in questo caso no, si è semplicemente andati a capo
- anche la titolazione dei paragrafi è titolo

Alcune norme

Titolo


COMPRA-VENDITA CASE
e TERRENI
L. 2,50 per parola
CASA moderna, rifiniture signorili, costruzione solidissima, paraggi Saragozza vendesi 230.000 trattabili. Scrivere CASSETTA 12 T UNIONE PUBBLICITA' ITALIANA, Bologna. 12379

- Anche le pubblicità hanno i loro titoli


ANNUNZI SANITARI

IL RESTO DEL CARLINO - 2 Novembre 1939 A. XVIII

- in caso di decorazioni tipografiche (come linee o decorazioni) cerchiamo di non includerle

Alcune norme

Testo

BOLLETTINO METEOROLOGICO

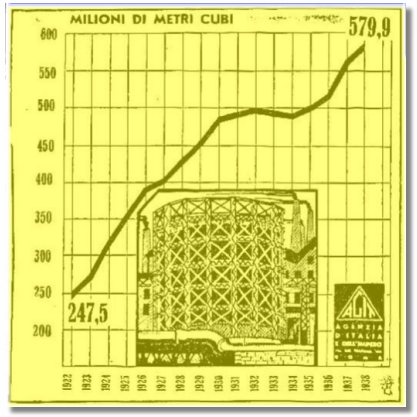
CITTA	Temperatura Massima	Temperatura Minima
Bologna	+ 8.7	+ 4.7
[illegible]	+ 15.0	+ 11.2
[illegible]	+ 10.8	+ 5.7
[illegible]	+ 11.2	+ 4.5
[illegible]	+ 14.6	+ 12.6
[illegible]	+ 17.4	+ 11.4
[illegible]	+ 8.0	+ 6.0
[illegible]	+ 12.8	+ 9.5
[illegible]	+ 8.3	+ 2.0
[illegible]	+ 10.0	+ 0.2
[illegible]	+ 12.8	+ 7.8
[illegible]	+ 13.2	+ 4.3
[illegible]	+ 15.0	+ 8.5
[illegible]	+ 13.0	+ 11.0
[illegible]	+ 16.5	+ 8.0
[illegible]	+ 16.5	+ 8.9
[illegible]	+ 17.4	+ 8.0
[illegible]	+ 17.8	+ 8.4
[illegible]	+ 17.0	+ 13.2
[illegible]	—	—
[illegible]	+ 18.6	+ 5.0
[illegible]	+ 17.0	+ 7.4
[illegible]	+ 17.5	+ 10.0
[illegible]	+ 22.2	+ 8.5
[illegible]	+ 20.2	+ 17.2
[illegible]	+ 14.1	+ 12.7

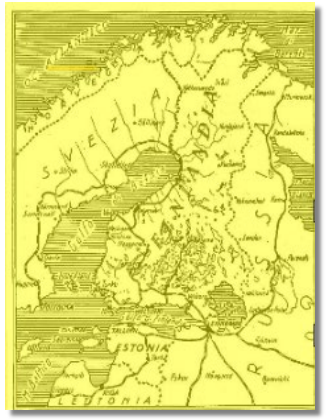
VENEZIA, 28 - PREVISIONI DEL TEMPO PER IL VENETO, L'EMILIA, LA ROMAGNA E LE MARCHE VALEVOLI FINO ALLE 19 DEL 29: Condizioni mediocri. Venti deboli occidentali. Cielo coperto. Nebbie e foschie diffuse. Mare poco mosso.

- le tabelle sono testo
- è buona norma selezionarle singolarmente separandole dal resto del testo

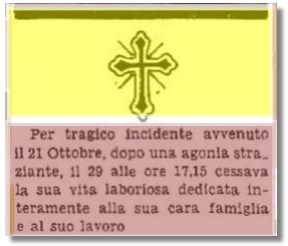
Alcune norme

Disegno

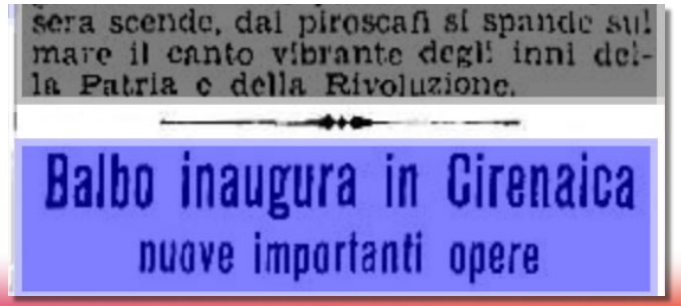




- Grafici e mappe sono considerati disegno
- se dovesse esser presente una didascalia va inclusa nell'area



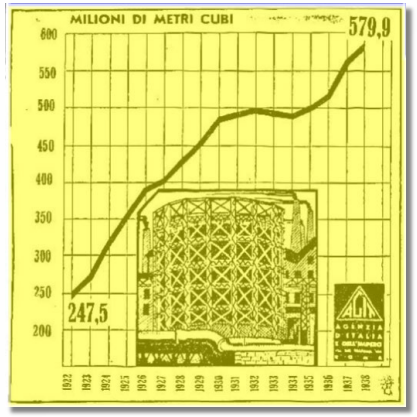
- anche marker tipografici come questo sono disegni

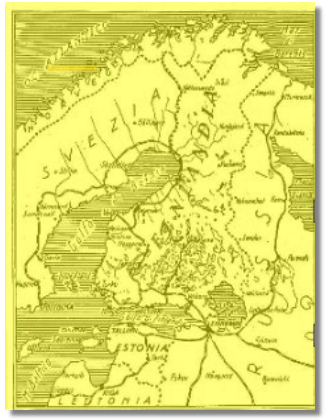


- se di questo tipo invece non vanno inclusi

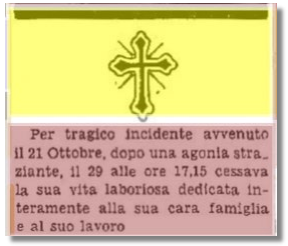
Alcune norme

Disegno

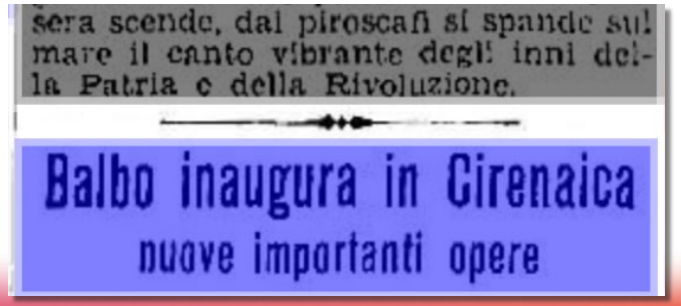




- Grafici e mappe sono considerati disegno
- se dovesse esser presente una didascalia va inclusa nell'area



- anche marker tipografici come questo sono disegni



- se di questo tipo invece non vanno inclusi

Alcune norme

Fotografia


Giuseppe Lombrassa
nominato Commissario
per le Migrazioni Interne

Uno scompartimento dotato di sei cuccette.

- Vanno incluse anche le didascalie
- nel caso ci sia un'immagine con un titolo, il titolo va inserito come titolo, e non incorporato nell'immagine

Alcune norme

Pubblicità

- Le singole pubblicità vanno separate
- Disegni e testi all'interno delle pubblicità non vanno ulteriormente separati

Alcune norme

Pubblicità

Spesso la categorizzazione pubblicità o testo non è immediata. La logica che si è scelto di seguire è la seguente: se è l'annuncio è stato pagato per esser pubblicato è da considerarsi pubblicità

In questo caso i cinema non hanno pagato per pubblicare, è quindi da considerarsi **testo**

Lavori o immobili

AVVISI D' INDOLE COMMERCIALE
L. 2.50 per parola

A pagamento mensile: Impermeabili, paltò, novità, stoffe. Gianni, D'Azeglio 46. 12524

ACQUISTIAMO autocarri, automobili, macchinario ogni specie fuori uso. Prezzi massimi. Pasquali, Oriani 40. Telefono 20983. Bologna. 12620

ADDIZIONATRICI, calcolatrici, macchine da scrivere. Vastissimo assortimento. Cambi. Fornisconsi rivenditori. UMA, Telefono 33-666, Parini 14 interno. 12315

APRILIA acquistata giugno carrozzeria speciale come nuova vendesi L. 31500. Scrivere CASSETTA 7 V UNIONE PUBBLICITA' ITALIANA, Bologna.

BIANCHI S 9 Cabriolet lussuoso perfetto adattissimo metano svendo 13.500. Acquisto Ford 3 litri oppure 522 purchè perfettissime. Cassetta 33822 Z Unione Pubblicità Italiana, Trieste.

CAMION Lancia 35 quintali 9750. Berline, Alfa, Balilla, 514, 509, vere occasioni. Martinelli, S. Giorgio 3. 12636

CAMIONISTI acquistate gassogeno « Dux » vera economia 0.80 carbonella sostituisce un litro carburante qualsiasi applicazione. Martinelli, S. Giorgio 3. 12638

FIAT 500 1100 1500 Lancia Aprilia non oltre otto mesi vita acquistansi. Indicare prezzo ristretto contanti. Scrivere CASSETTA 6 T UNIONE PUBBLICITA' ITALIANA, Bologna. 12564

Nuove nascite

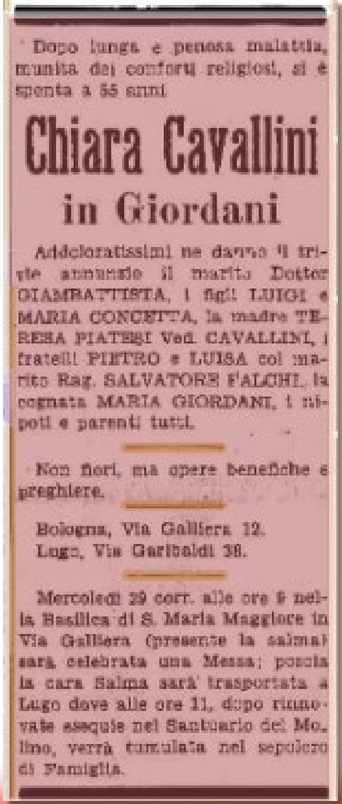
FIOCCHI BIANCHI

GIAN CARLO VACCARI annuncia con gioia la nascita del fratellino

GIAN FRANCO

Bologna, 22 novembre 1939-XVIII. Via Savioli 10.

Necrologi



Dopo lunga e penosa malattia, munita dei conforti religiosi, si è spenta a 55 anni

Chiara Cavallini in Giordani

Addoloratissimi ne danno il triste annunzio il marito Dottor GIAMBATTISTA, i figli LUIGI e MARIA CONCETTA, la madre TERESA PIATESI Ved. CAVALLINI, i fratelli PIETRO e LUISA col marito Rag. SALVATORE FALCHI, la cognata MARIA GIORDANI, i nipoti e parenti tutti.

Non fiori, ma opere benefiche e preghiere.

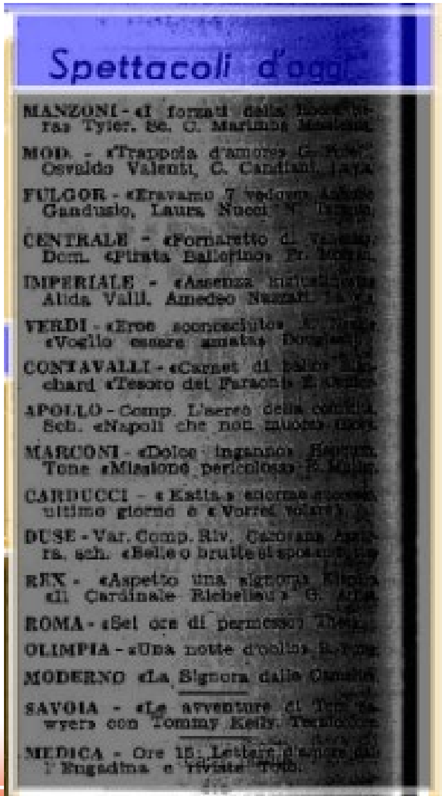
Bologna, Via Galliera 12.
Lugo, Via Garibaldi 38.

Mercoledì 29 corr. alle ore 9 nella Basilica di S. Maria Maggiore in Via Galliera (presente la salma) sarà celebrata una Messa; poscia la cara Salma sarà trasportata a Lugo dove alle ore 11, dopo rinnovate esequie nel Santuario del Molino, verrà tumulata nel sepolcro di Famiglia.

Film e spettacoli

TRAPPOLA D'AMORE
Oggi al MODERNISSIMO

E' un film spassoso diretto da ... tarazzo con Carla Candiani, Lia ... landini, Osvaldo Valenti, Giuseppe Porelli - Prima visione.



Spettacoli d'oggi

MANZONI - «I forzati della ... ras Tyler. Sc. C. Marinna ...
MOD. - «Trappola d'amore» ... Osvaldo Valenti, C. Candiani, ...
FULGOR - «Eravamo 7 vedove» ... Gandusio, Laura Nucci ...
CENTRALE - «Fornaretto di ... Dom. «Pirata Ballerino» ...
IMPERIALE - «Assenza ingiustificata» Alida Valli, Amedeo Nazzari ...
VERDI - «Eroe sconosciuto» ... «Voglio essere amata» ...
CONTAVALLI - «Carnet di ballo» ... chard «Tesoro del Faraone» ...
APOLLO - Comp. L'aereo della ... Sch. «Napoli che non muore» ...
MARCONI - «Dolce inganno» ... Tone «Missione pericolosa» ...
CARDUCCI - «Katia» enorme successo ultimo giorno e «Vorrei ...
DUSE - Var. Comp. Riv. ... ra, sch. «Belle o brutte si sposano ...
REX - «Aspetto una signora» ... di Cardinale Richelieu» ...
ROMA - «Sei ore di permesso» ...
OLIMPIA - «Una notte d'oblio» ...
MODERNO «La Signora dalle Camelie» ...
SAVOIA - «Le avventure di Tom Sawyers con Tommy Kelly ...
MEDICA - Ore 15: Lettere d'amore ... l'Engadina e rivista ...

Alcuni casi peculiari

PICCOLI AVVISI

MINIMO 10 PAROLE OGNI AVVISO

Si ricevono presso la

UNIONE PUBBLICITA' ITALIANA

N.B. - Tutti gli avvisi provenienti da agenzie sono soggetti alla tariffa « Commerciale ».

Alcuni casi peculiari


Anno LV N. 284 - Italia Impero Colonie, cent. 30

ABBONAMENTI

il Resto del Carlino

Bologna - Giovedì 30 Novembre 1939-XVIII

TARIFFA PER LE INSERZIONI

UNIONE PUBBLICITA' ITALIANA S. A.


il Resto del Carlino

ABBONAMENTI

per il 1940-XVIII-XIX

EDIZIONE DELLA SERA

il Resto del Carlino

ABBONAMENTI CUMULATIVI


Nel vostro esclusivo interesse citate sempre nei vostri ordini e offerte

"IL RESTO DEL CARLINO"

- Queste sono pubblicità in quanto riportano costi e tariffe per gli abbonamenti al giornale

Alcuni casi peculiari

STATO CIVILE DI BOLOGNA

Denuncie dal 25 Novembre 1939-XVIII

Nati ____ 6
Morti ____ 16
Matrimoni ____ ..

FIOCCHI BIANCHI

GIAN CARLO VACCARI annuncia con gioia la nascita del fratellino

GIAN FRANCO

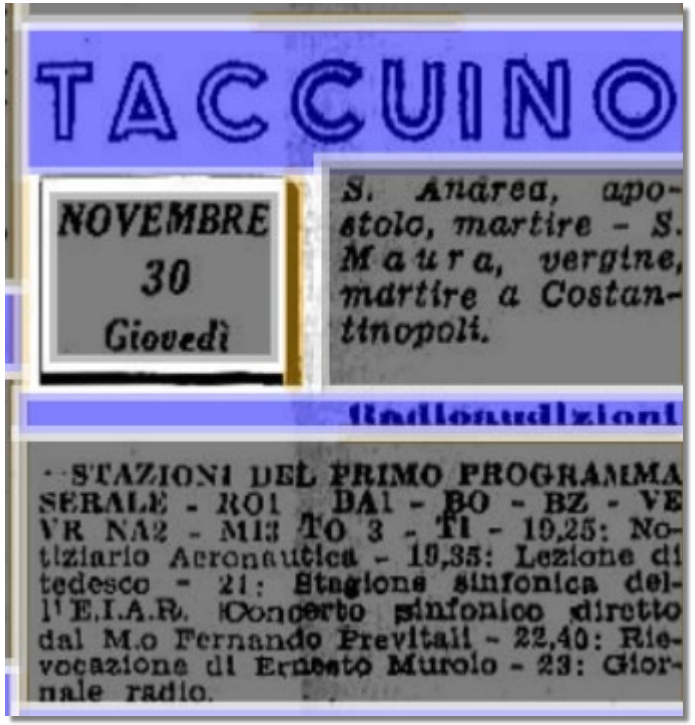
Bologna, 22 novembre 1939-XVIII.
Via Saviolì 10.

Ai nati annunciati in questa rubrica l'Unione Pubblicità Italiana regala un biglietto della Lotteria «E 42» col quale possono vincere molti milioni

Al piccolo Gian Franco Vaccari è stato assegnato il biglietto della Lotteria E. 42 N. 40952 Serie I.

- E’ un <u>testo</u> in quanto un dato sullo stato civile
- E’ una <u>pubblicità</u> in quanto annuncio pagato dalla famiglia
- E’ un <u>testo</u> in quanto riporta una notizia

Alcuni casi peculiari



TACCUINO

NOVEMBRE
30
Giovedì

S. Andrea, apostolo, martire - S. Maura, vergine, martire a Costantinopoli.

Radioaudizioni

STAZIONI DEL PRIMO PROGRAMMA SERALE - RO1 DA1 - BO - BZ - VE VR NA2 - MI3 TO 3 - TI - 19,25: Notiziario Aeronautica - 19,35: Lezione di tedesco - 21: Stagione sinfonica dell'E.I.A.R. Concerto sinfonico diretto dal M.o Fernando Previtali - 22,40: Rievocazione di Ernesto Murolo - 23: Giornale radio.

- La data è testo – da non includere il riquadro

Immagini e titolo

Nel caso di un'immagine con didascalia ma anche titolo, che non ha corrispondenza con altre aree di testo, il titolo non va incorporato nell'immagine ma marcato come titolo

Titoli e sottotitoli non riunibili

In casi come questo si è deciso per ora di lasciare la suddivisione delle bounding box

Casi da attenzionare

LA STAGIONE LIRICA AL COMUNALE

LA "CARMEN,, E LA GLORIA

Giorgio Bizet, nato a Parigi da famiglia di musicisti il 25 ottobre 1838, non ebbe vita troppo fortunata. Autore di molta musica da camera, orchestrale e per il teatro, conseguì sempre successi discreti, avvelenati dalla critica avversa, e non arrivò mai da vivo a conoscere un trionfo completo ed autentico. Anche la sua ultima opera, il capolavoro, la sua oggi gloriosa «Carmen», quando fu eseguita per la prima volta al Teatro dell'«Opéra Comique» di Parigi, la sera del 3 marzo 1875, fu accolta da quel pubblico con la più grande freddezza.

Tre mesi dopo, il 3 giugno dello stesso anno 1875, il Bizet moriva a Bougival quasi improvvisamente, a 37 anni, né mai si è potuto sapere con certezza per quale specie di malattia; opinando i più esser stata conseguenza di lunghi e persistenti attacchi di angina tonsillare sofferti durante la fanciullezza.

...lla gente; i due nuovi innamorati scambiandosi apertamente parole ...ore dolci e ardenti.

...camillo, poi, lascia Carmen per il ...compito, e le amiche avvertono ...men, che Don José, nascosto tra la ...alca, la sta spiando. Carmen, allora, si decide ad attendere, sola, il vecchio amante, mentre la folla si è già ...a riversata nel Circo. E qui si svolge un'azione fra le più violenti e drammatiche che immaginar si possa, creata dall'urto di due passioni indomabili. Lui, comparso, prima prega la ...na, poi la minaccia, lasciandosi sa... n poco a poco il sangue alla testa. Lei gli ride sulla faccia, lo sfida e ... ultimo si toglie dal dito un anello ...la fu donatole e lo getta lontano con disprezzo.

...utto è finito. Don José, accecato ...l'impeto del sangue, trae la lama spagnola e con quella trafigge la don-

Nel caso in cui la pagina sia stata acquisita leggermente storta non sarà possibile identificare correttamente le colonne, e il problema si porrà maggiormente nelle colonne a tutta altezza. Non vi è soluzione (al momento) per questa problematica.

Infatti:

- Layout Parser lavora solo su aree rettangolari, non è possibile creare trapezi per seguire la colonne non perfettamente verticali
- Suddividere la colonna in n rettangoli successivi andrebbe a confondere il modello

Soluzione: *fai come faresti** e lascia un flag in maniera tale che la pagina possa esser facilmente riconoscibile come «problematica»

* cerca di trovare il miglior compromesso tra aree incluse e aree sovrapposte

Questioni aperte

- Il sistema riconosce anche le gerarchie Titolo – Articolo?
  - Se si, lo stiamo allenando anche per questo?
    - Se no, dovremmo ri allenarlo?
  - O servono ulteriori motori (ancora da identificare)?

  - [SIL] Layout Parser non crea le gerarchie
    la nostra intenzione sarebbe di estrapolarla a valle
    con del codice nostro ad hoc

Grazie dell'attenzione

http://visitlab.cineca.it/
visitlab@cineca.it