Corriere della Sera Digital Project

Anno I. Domenica-Lunedì, 5-6 Marzo 1876 Num.

CORRIERE DELLA SERA

Prezzi d'associazione

Centesimi 5 in Milano — 7 fuori — numeri arretrati 10.

Ufficii di Direzione ed Amministrazione in Milano, Galleria Vittorio Emanuele, 77

Prezzo delle inserzioni

Abbonamento straordinario

Un abbonamento straordinario è aperto tutto l'anno 1870 ai seguenti prezzi:

Per Milano a domicilio.
Dal 5 marzo al 31 dicembre 1876 L. 10.

Per tutto il Regno, franco di porto.
Dal 5 marzo al 31 dicembre 1876 L. 15.

Estero (Stati dell'Unione Postale).
Dal 5 marzo al 31 dicembre 1876 L. 25.

MILANO, 5 MARZO

nificazione d'Italia, ha restaurato le finanze. Se domani dovesse abdicare, potrebbe, con l'orgoglio che dà l'adempimento d'un gran compito, esclamare: *Nunc dimittis, domine.* Da un disavanzo annuo spaventevole ci ha condotti al pareggio. Non ancora, dite? Ebbene, sia: mancano venti, mancano trenta milioni: che sono appetto ai 700 che mancavano dieci anni fa? Qualchecosa di peggio che le finanze turche. Allora si discuteva sul fallimento dello Stato, e si cercava di

fanatiche delle campagne o delle plebi voltabili e nervose delle città.

><

Questo giornale, che è moderato, e vuol essere lo specchio fedele dei pensieri di chi scrive, e delle persone savie che vorranno aiutarci de' loro consigli, — e li invochiamo, giacchè, se siamo indipendenti, non vogliamo restare isolati, — non promette di essere di più facile contentatura dell'altra gente del suo partito; e però non si farà scrupolo di

piace ricordarci che tu, pubblico, non t'interessi che mediocremente ai nostri odî ed ai nostri amori; che vuoi anzitutto essere informato con esattezza; ci place serbare, di fronte a' nostri amici migliori, la nostra libertà di giudizio, ed anche, se vuolsi, quel diritto di *frondismo* ch'è il sale del giornalismo.

><

Sentiamo dire ancora: Badate, voi dividete il partito. — Davvero? ma era forse diviso il partito quando esistova a

intiepidita una sola delle numerose s... che ci hanno spontaneamente circon... questi giorni: non hanno servito a... a riscaldarlo. La nascita del *Corriere* ... è pura come un raggio di sol... tutti è nota la privata iniziativa d... figlio. Per vivere non ha bisogno di... narsi a chi comanda, come non ha ... di blandire le passioni ed i pregiudizi ... della. *Fallerus quia externus* si dice. Il *Corriere della sera* può essere ... rente alla calunnia, e sorridere ... non è destinato a morire domani. ... nire è suo.

IFLA International Newspaper Conference
Bibliothèque Nationale de France - 11-13 April 2012

- **Claudio Albanese, IDM, *(Italy)***
- **Walter Colombo, Corriere della Sera, *(Italy)***
- **Shalev Vayness, ISAKO, *(France)***

Isako
A passion for innovation

RCS MEDIAGROUP

IDM Group
integra document management

Corriere della Sera Digital Project

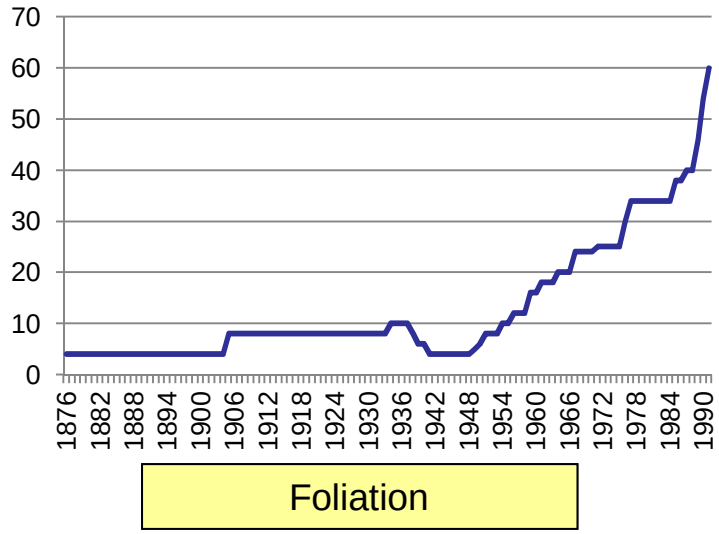
Corriere della Sera in figures

About Corriere della Sera

- First published in Milano in 1876
- With a circulation of 15.000 copies
- In 1920 circulation of 600.000 copies

Today

- Daily circulation of 480.000
- 15 Regional editions
- 150 pages daily in national and local editions.



Multiple Daily Editions

- Starting in 1883, the newspaper published two afternoon editions.
- From 1890 it was published in three editions.
- From 1903 four editions, two in the morning and two in the afternoon

Dacs Project in figures

- **2.000.000 pages**
- Up to 12.000.000 articles
- Up to 60.000.000 digital objects

The Historical Archive

- The entire collection in paper is bound in volumes divided by edition.

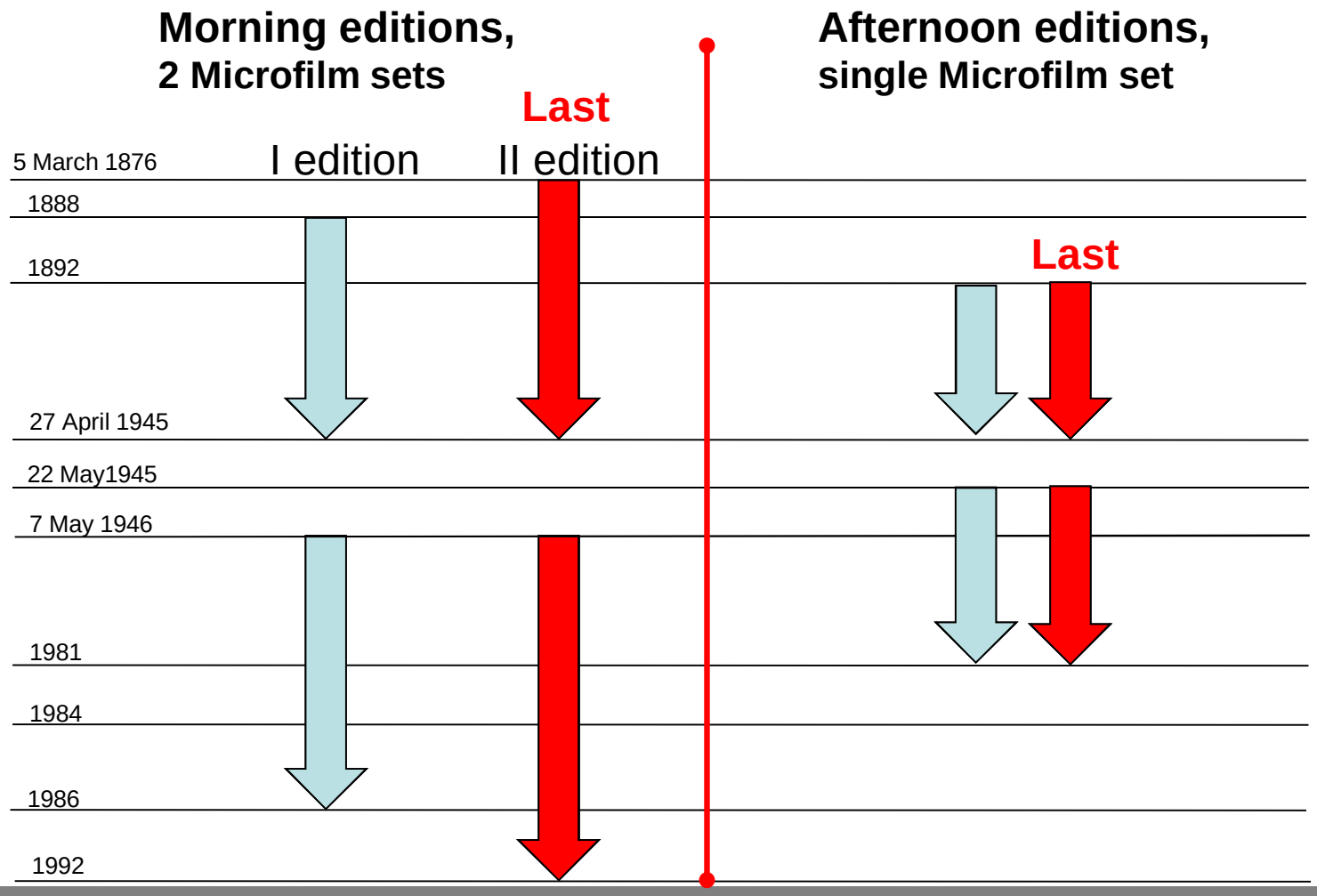
- In 1970 started the creation of microfilm, three sets of microfilm, each one with master and copy:
  - two sets for the morning editions (Last Edition, Previous Editions),
  - one set for the afternoon editions.

- In 1992 initiated the digital copy, with the full text of all published articles.

Corriere della Sera Digital Project

Microfilm and Editions

Daily Editions and respective microfilm sets



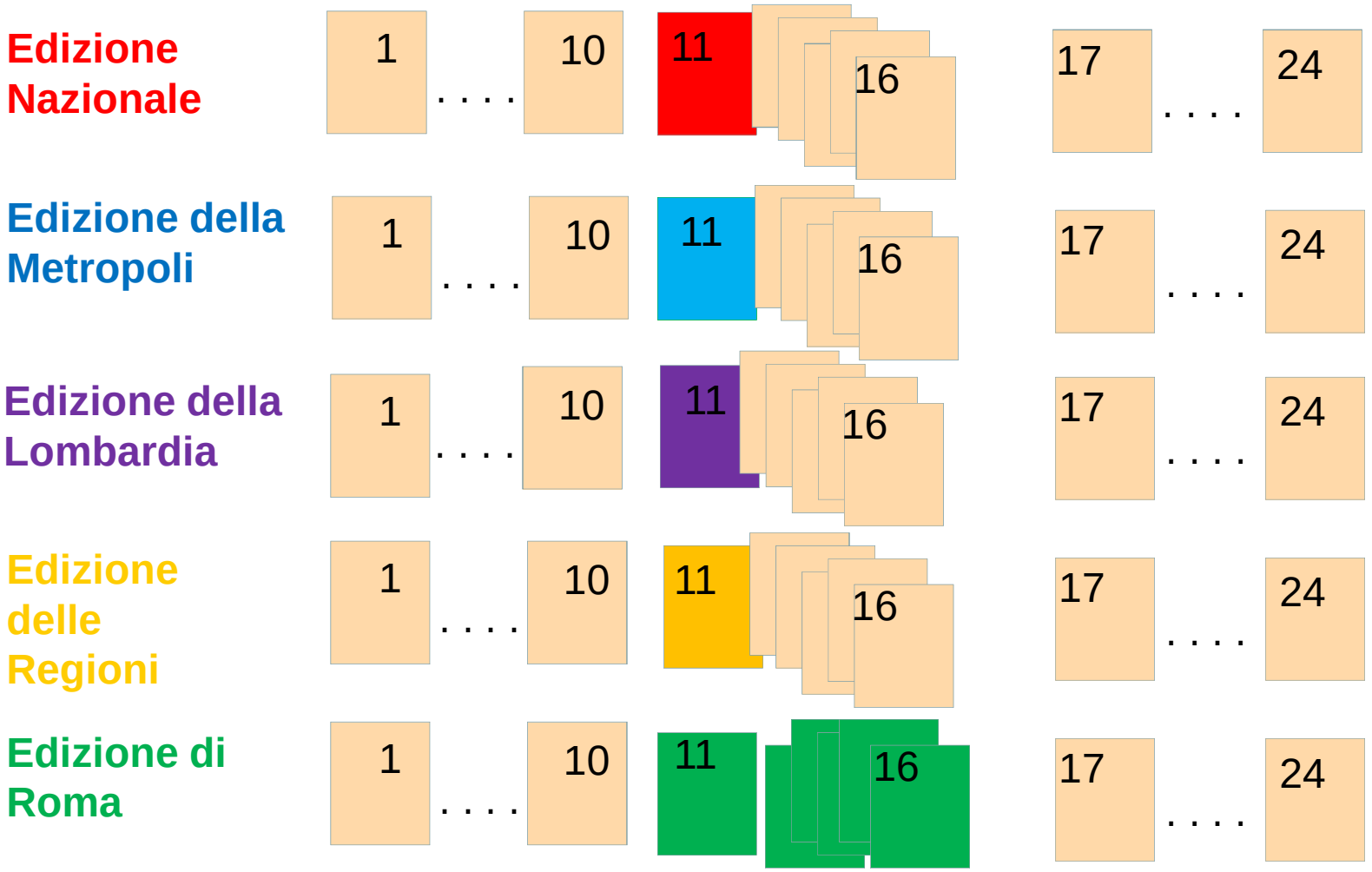
Isako
A passion for innovation

RCS MEDIAGROUP

IDM Group
integra document management

Corriere della Sera Digital Project

Geographical Editions



Isako
A passion for innovation

RCS MEDIAGROUP

IDM Group
integra document management

La Fondazione Corriere della Sera

- Founded in 2001, to preserve and spread the cultural heritage of Corriere della Sera.
- Develops, with research and publications, conferences and exhibitions, the knowledge of Corriere and of all editorial World of RCS.
- A particular focus on the conservation of the historical paper archives.

Objectives, Vendor Selection & Pre-Production Steps

Objectives

- Paper conservation
- Allow better access to the archives
- Increase readership of Corriere - business opportunities.

Phase 1

- Vendor selection
  - Short list of 4 vendors
  - Identify the correct balance between cost and quality :
    Text accuracy : 100% for titles, 95% body of articles.
- Define deliverables
- Define Service Level (SLA)
- Value the need of following market standards
- **Determine scanning source, paper vs microfilm.**

IDM Group
integra document management

- A Global Document Process Outsourcer.
- Setting up processes, organizing workloads and providing skilled resources for the CorrSera digital project.

Isako
A passion for innovation

- Experienced provider of digitization workflow solutions, capable of adapting its existing solutions to CorrSera digital project's specific needs.
- Advising and assisting CS and IDM in making key technical choices such as data formats.

Deliverables & Quality Level

- **A workflow system is fundamental for the management and control of the project**
  - Volume of digital objects
  - The complexity of the data
  - The need to manage multiple daily editions, both in time and in location.

- **Dealing with multiple daily editions**
  - Unique Last Edition
  - Multiple Previous Editions
  - **Different deliverables for “Last” and “Previous” editions.**

- **Deliverables**
  - Images : Pages and Articles in various formats and resolutions.
  - PDF : With Hidden Searchable text.
  - ALTO / METS and an Editorial XML.

Corriere della Sera Digital Project

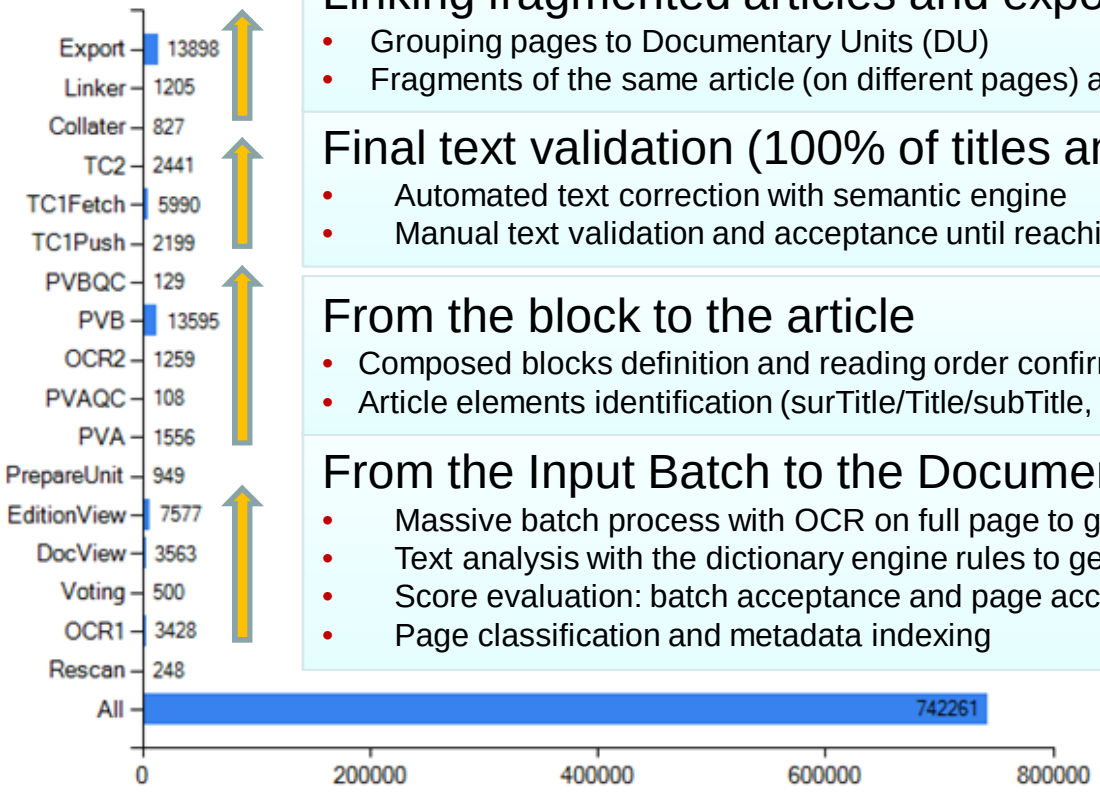
Production Workflow

NCS - DACS v1220

Home | Batches | Units | Pages | Administration

Pages overview:

All	Rescan	OCR1	Voting	DocView	EditionView	PrepareUnit	PVA	PVAQC	OCR2	PVB	PVBQC	TC1Push	TC1Fetch	TC2	Collater	Linker	Export
742261	248	3428	500	3563	7577	949	1556	108	1259	13595	129	2199	5990	2441	827	1205	13898



Linking fragmented articles and export of deliverables

- Grouping pages to Documentary Units (DU)
- Fragments of the same article (on different pages) are linked together

Final text validation (100% of titles and 95% of text)

- Automated text correction with semantic engine
- Manual text validation and acceptance until reaching the quality target

From the block to the article

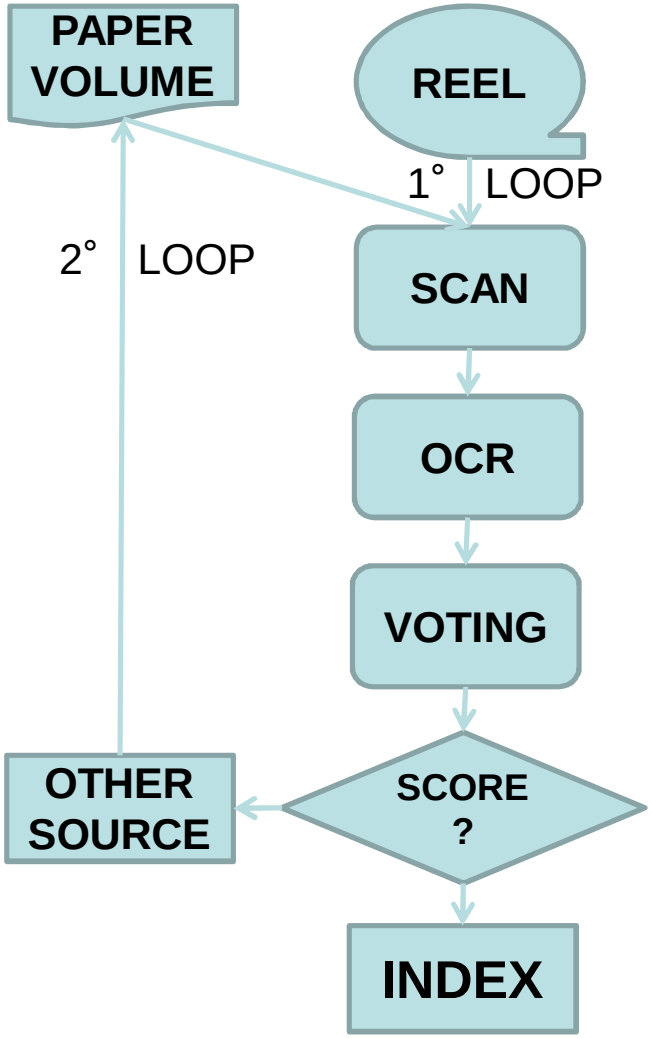
- Composed blocks definition and reading order confirmation
- Article elements identification (surTitle/Title/subTitle, text, photo, author)

From the Input Batch to the Documentary Unit (DU) & Edition

- Massive batch process with OCR on full page to get the OCR score (characters)
- Text analysis with the dictionary engine rules to get the dictionary score (words)
- Score evaluation: batch acceptance and page acceptance
- Page classification and metadata indexing

Corriere della Sera Digital Project

Challenge 1 : *Scan and Image Quality*



Low quality input will produce poor results:

Wasted Effort, Unusable Outcome !

MICRO vs PAPER

- MICRO to scan is a **direct copy** of MASTER
- PAPER for selected collection.

Initial OCR on all pages

- Without “preventive” selection (noise included)
- OCR score per page and batch average
- DICTIONARY score per page and batch average.

QC is the first human activity

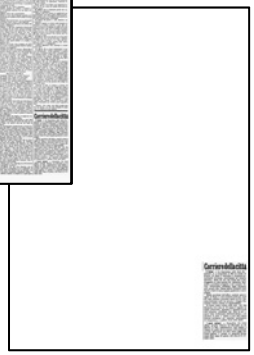
- Rejection of single page or of the complete batch
- The accepted pages are indexed with all metadata
- Pages go forward within the workflow.

Data Formats and Deliverables

- Adding an Editorial XML (NITF) to the METS / ALTO.
  - Cross-Referencing the NITF with the METS/ALTO.
  - Adding semantic and editorial information to the ALTO so that the NITF can be fully derived from the METS / ALTO.

- Linking distinct physical fragments of the same article into a **single logical article in the METS & NITF**.

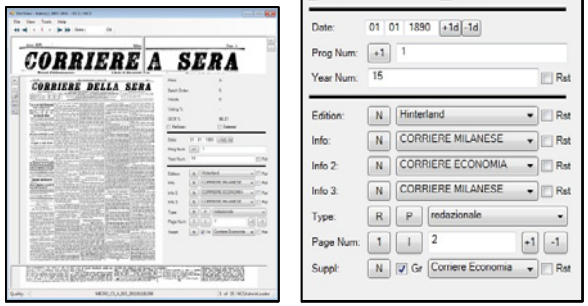
Article Images are kept within the physical page.



```
<mets:div ID="divarticle15" TYPE="article" DMDID="CS_18820304_L_art0000015">
  <mets:fptr FILEID="CS_18820304_L_art0000015_nitf_xml" />
  <mets:div ID="divarticle15-1" TYPE="article-part" ORDER="1" >
  <mets:div ID="divarticle15-2" TYPE="article-part" ORDER="2" >
</mets:div>
```

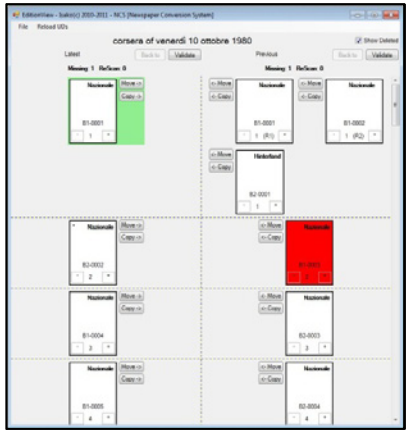
```
<body class="Article">
  <body.head>
    <hedline>
      <hl1 class="Title">Corriere della Città</hl1>
    </hedline>
  </body.head>
  <body.content id="p1.ART_000010">
    <block class="Physical" id="p1.PB000025">
    <block class="Physical" id="p1.PB000026">
  </body.content>
  <body.content id="p2.ART_000001">
    <block class="Physical" id="p2.PB000001">
    <block class="Physical" id="p2.PB000002">
  </body.content>
  <body.end />
</body>
</nitf>
```

Optimized and Dedicated Tools



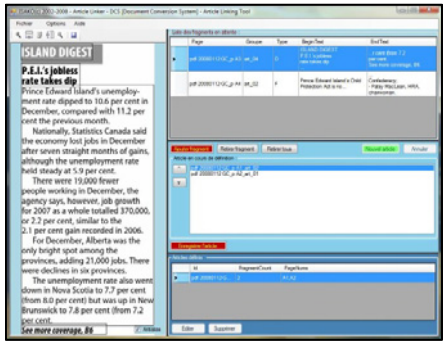
- **DocView**
  Set and control the page's "objective" metadata.

- **EditionView**
  Manage multiple DUs, multiple editions and multiple occurrences of pages for a given date.



Full clipping of “Last Edition” Units including

- Tagging specific blocks
  - Occhiello, Title, subTitle, Author, Photo, Legend.
- Specific semantic tagging : Obituaries.

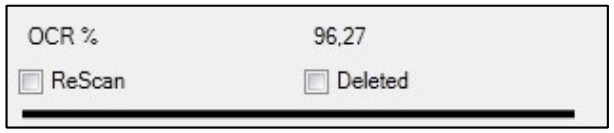


Solving the productivity challenge

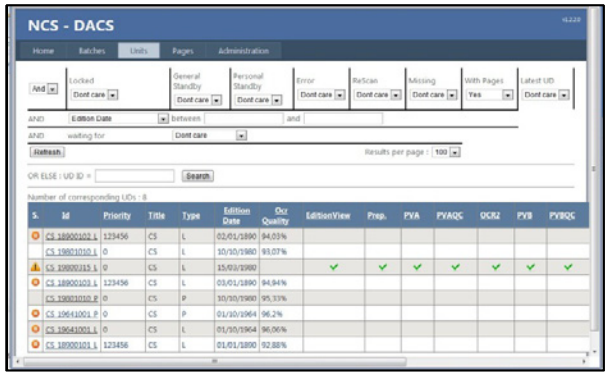
- Clipping : **Dedicated 2 step process** including adapting the standard PageView tool for each step.
- Linking : **Dedicated tool** to combining several physical fragments into a single logical article.

Distributed “Just-in-Time” Quality Control

- DocView is also the QC of image quality (rescan).

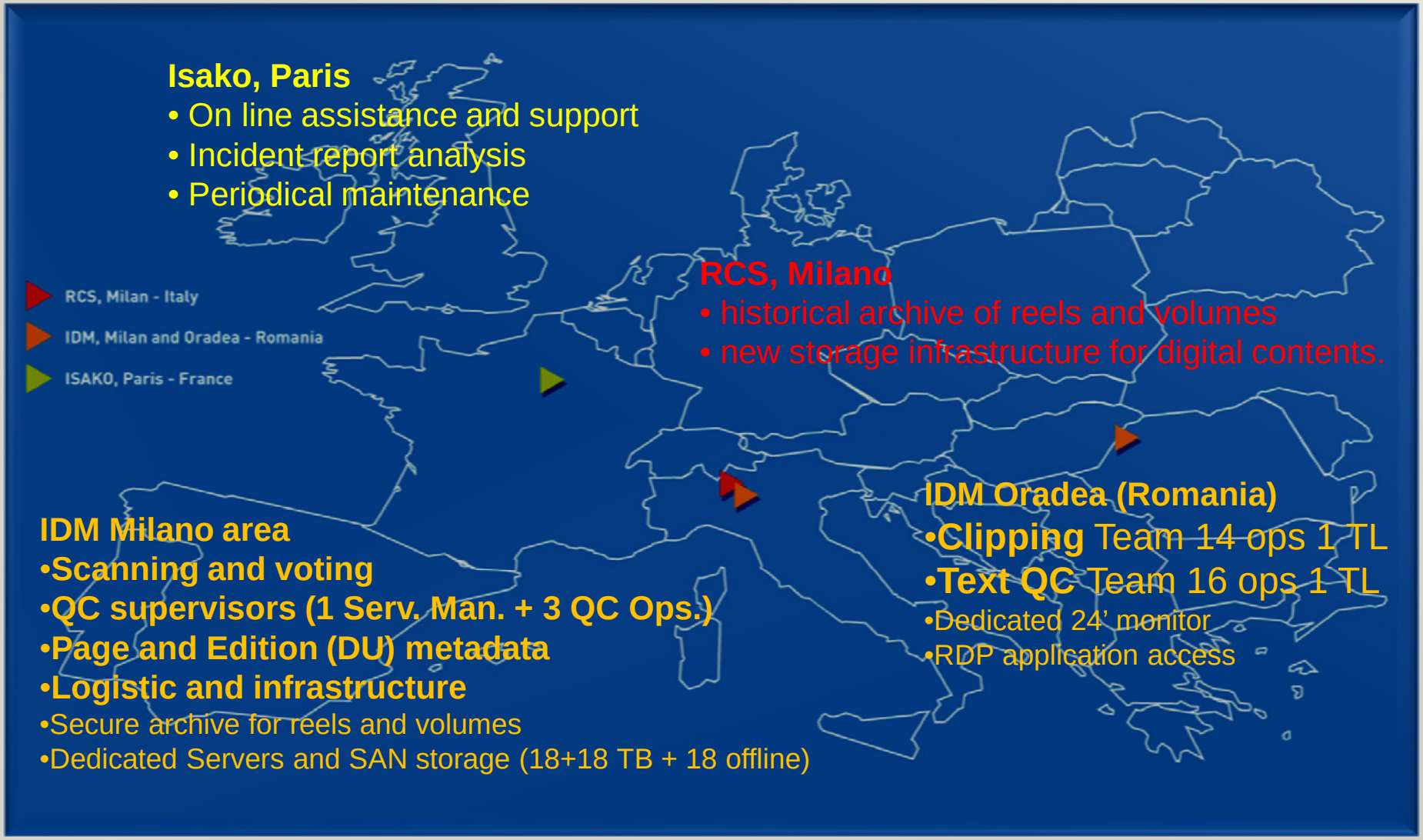


- **Double Internal Quality Control of the clipping**
  - Controlling by an experienced operator or a team-leader.
- **Built-in Stand-by mechanism.**
  - Accessible by CorrSer if necessary.
- **Ad-Hoc QC done by CorrSer inside the actual “live” process.**



Corriere della Sera Digital Project

Challenge 5 : *Multiple Geographical Sites*



Isako, Paris
- On line assistance and support
- Incident report analysis
- Periodical maintenance

RCS, Milano
- historical archive of reels and volumes
- new storage infrastructure for digital contents.

IDM Milano area
- **Scanning and voting**
- **QC supervisors (1 Serv. Man. + 3 QC Ops.)**
- **Page and Edition (DU) metadata**
- **Logistic and infrastructure**
- Secure archive for reels and volumes
- Dedicated Servers and SAN storage (18+18 TB + 18 offline)

IDM Oradea (Romania)
- **Clipping** Team 14 ops 1 TL
- **Text QC** Team 16 ops 1 TL
- Dedicated 24' monitor
- RDP application access

Isako
A passion for innovation

RCS MEDIAGROUP

IDM Group
integra document management

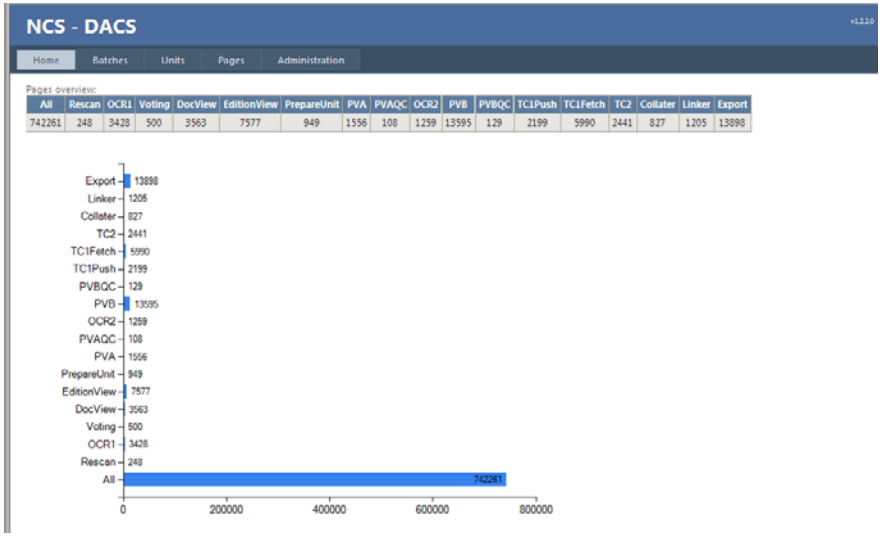
Corriere della Sera Digital Project

Challenge 6 : Production Management

Daily report

2012	PVA	PVB	TC2			AVG		OCR1	Voting	DV	EV	PU	OCR2	Push	Fetch	Col	LNK	Exp	94	245	AVG
19/03/2012	5000	7307	2133	0	1		19/03/2012	512	1122	7889	10817	7061	8110	5317	3191	6921	6301	5495	1	1	
20/03/2012	5930	5693	3076	1	1		20/03/2012	2015	1847	9874	5248	3135	6358	7884	8019	9262	9627	4058	1	1	
21/03/2012	4423	4844	3409	1	1		21/03/2012	3782	3477	0	10380	7584	6613	4059	4152	9162	8787	4486	1	1	
22/03/2012	2852	4251	2438	0	1		22/03/2012	2007	1037	4445	1284	1856	4154	4718	4623	7858	7579	2608	1	1	
23/03/2012	2666	3917	3000	1	1		23/03/2012	3432	3752	2290	5012	4061	4770	391	4491	6565	6311	5393	1	1	
24/03/2012	0	0	23			2.811	24/03/2012	4685	4821	0	0	0	0	4819	885	696	696	5402	1	1	4.698
25/03/2012	0	0	482				25/03/2012	2930	2137	0	0	0	0	0	4140	1171	1171	4092	1	1	

Real-Time Workload Control



Detailed reports allow for detecting and solving bottlenecks by immediate reallocation of resources.

- Step by step workload reports
- Batch, DU & page level analysis
- Administrative tools
  - Reports
  - Management scripts
  - Setup parameters
  - Error control

Isako
A passion for innovation

RCS MEDIAGROUP

IDM Group
integra document management

Delivery and Final Quality Control

- Automated controls
  - File formats
  - Image file resolution
  - Completeness of data sets
- Manual controls
  - Text quality
  - Image quality
- How to survive quality control

- **Avoid underestimating the logistics of the 'page by page' rescan operations**
  - Different possible input source retrieval
  - Time delay on DU completion & export.
- **Simplify rules of clipping (continuous training on the job)**
  - Questions are posted on the page and shared on line .
  - Common access via NCS between IDM & RCS.
- **Integrate a communication tool inside the workflow**
  - Stand-by logic from Operators to Team Leader
  - Collaborative annotations on batch/DU/page.