Corriere della Sera Digital Project

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)
Corriere della Sera
Digital Project

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 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

Daily Editions and respective microfilm sets

Morning editions, 2 Microfilm sets
- 5 March 1876
- 1888
- 1892
- 27 April 1945
- 22 May 1945
- 7 May 1946
- 1981
- 1984
- 1986
- 1992

Afternoon editions, single Microfilm set
- Last

Isako
A passion for innovation

RCS MediaGroup
Integra document management

Microfilm and Editions
<table>
<thead>
<tr>
<th>Edizione Nazionale</th>
<th>Edizione della Metropoli</th>
<th>Edizione della Lombardia</th>
<th>Edizione delle Regioni</th>
<th>Edizione di Roma</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>11</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>
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

- 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.
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.
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.
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

From the block to the article
• Composed blocks definition and reading order confirmation
• Article elements identification (surTitle/Title/subTitle, text, photo, author)

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

Linking fragmented articles and export of deliverables
• Grouping pages to Documentary Units (DU)
• Fragments of the same article (on different pages) are linked together

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
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.

Low quality input will produce poor results:
Wasted Effort, Unusable Outcome!
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.
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**

**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

**Challenge 5: Multiple Geographical Sites**
### Daily report

<table>
<thead>
<tr>
<th>Date</th>
<th>PV A</th>
<th>PV B</th>
<th>TC 2</th>
<th>AVG</th>
<th>OCR 1</th>
<th>OCR 2</th>
<th>DV</th>
<th>EV</th>
<th>PU</th>
<th>OCR 2</th>
<th>Push</th>
<th>Fetch</th>
<th>Col</th>
<th>LN K</th>
<th>Exp</th>
<th>94</th>
<th>24</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>19/03/2012</td>
<td>5000</td>
<td>7307</td>
<td>2133</td>
<td>0</td>
<td>1</td>
<td></td>
<td>7889</td>
<td>10817</td>
<td>7061</td>
<td>8110</td>
<td>5317</td>
<td>3191</td>
<td>6921</td>
<td>6301</td>
<td>5495</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20/03/2012</td>
<td>5930</td>
<td>5693</td>
<td>3076</td>
<td>1</td>
<td>1</td>
<td></td>
<td>9874</td>
<td>5248</td>
<td>3135</td>
<td>6358</td>
<td>7884</td>
<td>8019</td>
<td>9262</td>
<td>9627</td>
<td>4058</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21/03/2012</td>
<td>4423</td>
<td>4844</td>
<td>3409</td>
<td>1</td>
<td>1</td>
<td></td>
<td>3782</td>
<td>3477</td>
<td>0</td>
<td>10380</td>
<td>7584</td>
<td>6613</td>
<td>4059</td>
<td>4152</td>
<td>9162</td>
<td>8787</td>
<td>4486</td>
<td>1</td>
</tr>
<tr>
<td>22/03/2012</td>
<td>2852</td>
<td>4251</td>
<td>2438</td>
<td>0</td>
<td>1</td>
<td></td>
<td>2007</td>
<td>1037</td>
<td>4445</td>
<td>1284</td>
<td>1856</td>
<td>4154</td>
<td>4718</td>
<td>4623</td>
<td>7858</td>
<td>7579</td>
<td>2608</td>
<td>1</td>
</tr>
<tr>
<td>23/03/2012</td>
<td>2666</td>
<td>3917</td>
<td>3000</td>
<td>1</td>
<td>1</td>
<td></td>
<td>3432</td>
<td>3752</td>
<td>2290</td>
<td>5012</td>
<td>4061</td>
<td>4770</td>
<td>391</td>
<td>4491</td>
<td>6565</td>
<td>6311</td>
<td>5393</td>
<td>1</td>
</tr>
<tr>
<td>24/03/2012</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>2.811</td>
<td>24/03/2012</td>
<td>4685</td>
<td>4821</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4819</td>
<td>885</td>
<td>696</td>
<td>696</td>
<td>5402</td>
<td>1</td>
<td>1</td>
<td>4.698</td>
</tr>
<tr>
<td>25/03/2012</td>
<td>0</td>
<td>0</td>
<td>482</td>
<td>0</td>
<td>25/03/2012</td>
<td>2930</td>
<td>2137</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4140</td>
<td>1171</td>
<td>1171</td>
<td>4092</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

### 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 Mediac 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 online.
- 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.