OCTOBER 25, 2021

NEWSPAPER ARCHIVING SOLUTION
ANNEXURE

Table of Contents

Annexure-1

1 Solution Overview

The solution should provide an easy, efficient, and quality newspaper scanning system with storing and backup capabilities. The scanned text should be readable, and images should have the same quality as the actual copy. Storage and backup system should be able to handle the load. The platform's user interface should be easy, user-friendly, and understandable with the features of viewing, searching, and retrieving necessary information & views from the archived newspaper's data.

The proposed solution needs to ensure that the stored archive remain unchanged forever. Thus, the solution needs to use Write Once Read Many objectives. A complete data security plan needs to implement to protect archived data. The solution also has to have the capability to ingest both structured and unstructured data. Archived data need to be easy to search for the time needed. A cost-effective storage tier or separate storage can be in the plan with expandability option to grow over the period of time to match data growth.

This solution will include:

1. The application for processing, storing, searching and updating the news archive.
2. High available server for processing, viewing and updating the archive
3. A high available storage system appropriate for storing the high volume of data
4. A backup solution that will ensure the security of the data focusing on availability prospective
5. Provide network infrastructure for handling such a high volume of data
6. Provide security infrastructure for handling such a high volume of data
7. Provide power backup solution for the provided solution

The proposed solution does not need to include:

1. Scanning solution (scanning equipment and computing devices)
2. Scanning service

Annexure-2

2 Input

All the inputs in the storage and backup system should meet the specification for scanning and categorization of the scanned data. Specification for scanning and categorization is given below

2.1 Scanning Specification

2.1.1 Scan quality PPI:

The scanning process should be capable of producing 300 PPI images in different formats. The quality or "resolution" of the pictures in the newspapers should be between 300 PPI to 600 PPI based on the source quality. The below figure shows how the quality of the image depends on the resolution.

Low-res images can look blocky, pixelated or blurry on screen and in print, like this:

Fig: PPI deference

2.1.2 Storage format:

The scanning procedure may have many storage formats for scanned data. In this solution, we expect the format should be any from the below list depending on the text and black and white image quality:

1. 1-bit monochrome
2. 4-bit grayscale
3. 6-bit grayscale
4. 8-bit grayscale

In the case of colour pictures, the format should be any from the below list depending on the image quality:

1. 16-bit RGB tiff
2. 16-bit RGB jpg
3. pdf
4. pdf (OCR)

2.1.3 Approximate file size:

The file size is estimated to be different for each page depending on the quality of the picture and the PPI. The table below shows an approximate file size for A1 pages with 400 PPI.

Sample data				
dpi	jpg(kb)	tiff(kb)	pdf(kb)	OCR pdf(kb)
200	5239	5016	5244	-
300	10417	9963	10422	-
400	16858	15752	16863	-
OCR sample 300 dpi/ single page	-	-	3963	5016

The storage capacity required for the process depends on the total newspaper count, the number of pages the newspaper has, and the scanned image format. Assuming the newspaper count, the page count, and fixing the DPI to an average of 400 DPI, the table below shows the storage capacity required in this process:

Assumptions 1					
Maximum storage requires for 1 A1 page	dpi	jpg (MB)	tiff (MB)	pdf (MB)	OCR pdf (MB)
	400	20	20	20	30
for any single format(jpg/tiff/pdf) + OCR pdf total maximum storage for 1 A1 page	=	50 MB			

Assumptions 2						
target year	newspaper scan target number	total A1 page for single newspaper	total page after both side count	total storage required for a single newspaper as per assumption 1(MB)	total storage for newspaper scan target number each day (MB)	total storage for target year (TB)
3	10	6	12	600	6000	7

Based on the above assumptions it's clear that storage required for archiving 10 newspaper for 3 years should not be more than 7 terabytes.

2.1.4 Post scanning modification:

For scanning an A1 size paper, the scanned image should be A1 size document, where the page should be splinted into two.

2.2 Specification for Categorization

2.2.1 Name of the paper:

The newspaper's name in digital format should be tagged with the scanned data for each newspaper page in the database.

2.2.2 Date of the paper:

Bangla, English, and Arabic date of the newspaper in digital format should be tagged with the scanned data for each newspaper page in the database.

2.2.3 Edition of the paper:

Edition of the newspaper in digital format should be tagged with the scanned data for each newspaper page in the database.

2.2.4 Location of the paper:

The location (region) of the newspaper in digital format should be tagged with the scanned data for each newspaper page in the database.

2.2.5 Page number:

Page number of the newspaper in digital format should be tagged with the scanned data for each newspaper page in the database.

2.3 Meta-data diagram

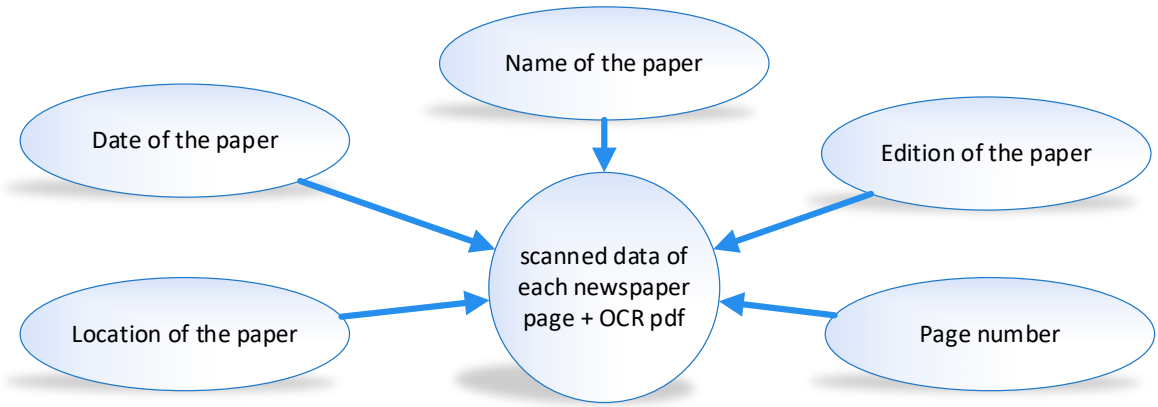

Fig: categorized data linked with scanned data

Annexure-3

3 Archiver Interface

The archiver interface should be user-friendly, flexible, error-free, and the functionality should meet the below specification:

3.1 Upload:

The interface should have the functionality where the archiver can upload scanned data tagged with all the meta-data. Also, each upload entry (scanned + meta-data) should have an auto-generated unique ID, upload date, uploaders name/ID, last modification date, and last modifiers name.

The interface should have a Bulk upload feature as per folder and file fixed format of naming convention.

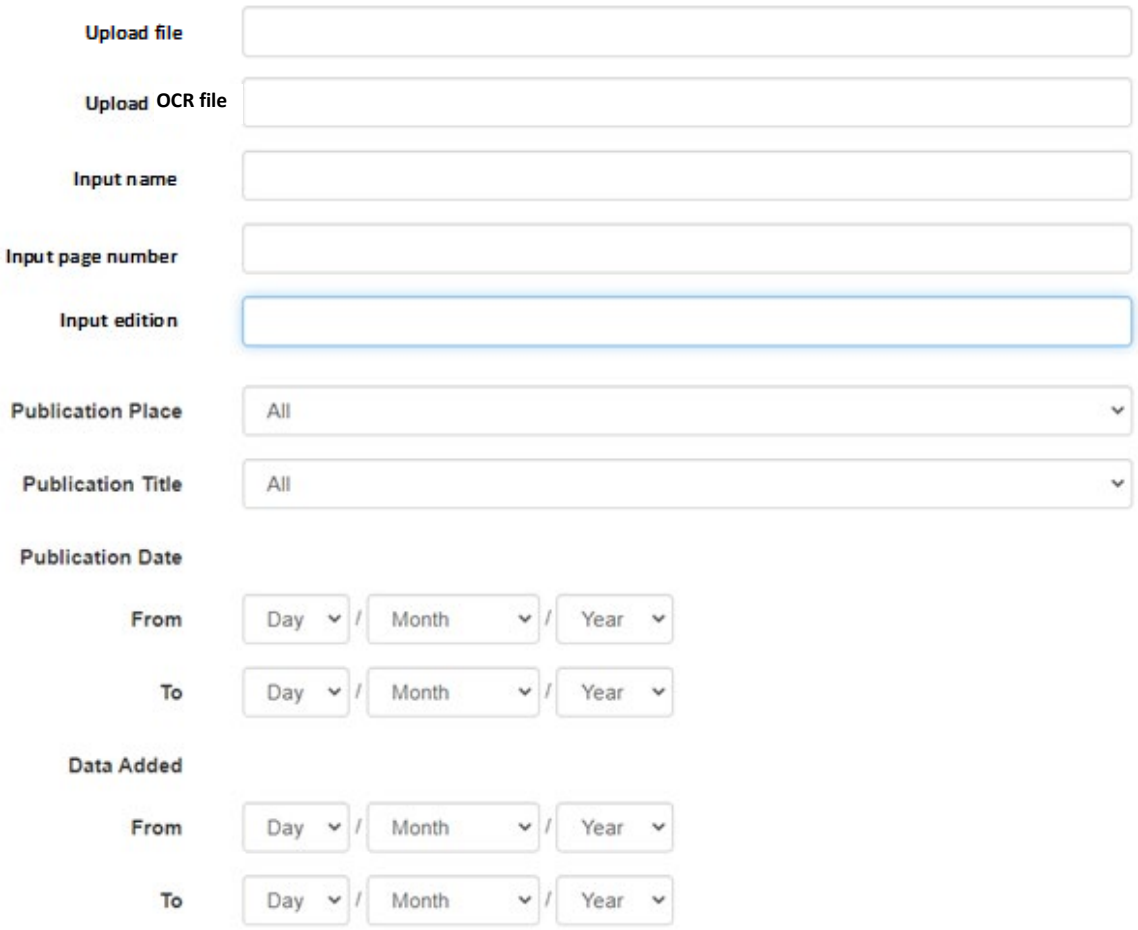

Fig: Archiver upload interface

3.2 Archiver View:

All the archivers should be able to view all the uploaded data, including all the meta-data in a list, and also, the archivers should be able to sort and filter through the list. Finally, by clicking on any entry should enable them to view as well as edit uploaded entries. The uploaded data list should be like the list given below:

ID (unique)	Name	Date	Page	Location	Edition	Upload date	Uploader's name/ID	Modification date	modifiers name
763eg32	Prothomalo	1/1/2021	1	Dhaka	1st	1/1/2021	archiver 1	1/1/2021	archiver 31
763eg33	Prothomalo	1/2/2021	2	Dhaka	2nd	1/2/2021	archiver 2	1/2/2021	archiver 32
763eg34	Prothomalo	1/3/2021	3	Dhaka	3rd	1/3/2021	archiver 3	1/3/2021	archiver 33
763eg35	Prothomalo	1/4/2021	4	Dhaka	4th	1/4/2021	archiver 4	1/4/2021	archiver 34
763eg36	Prothomalo	1/5/2021	5	Dhaka	5th	1/5/2021	archiver 5	1/5/2021	archiver 35
763eg37	Prothomalo	1/6/2021	6	Dhaka	6th	1/6/2021	archiver 6	1/6/2021	archiver 36
763eg38	Prothomalo	1/7/2021	7	Dhaka	7th	1/7/2021	archiver 7	1/7/2021	archiver 37
763eg39	Prothomalo	1/8/2021	8	Dhaka	8th	1/8/2021	archiver 8	1/8/2021	archiver 38
763eg40	Prothomalo	1/9/2021	9	Dhaka	9th	1/9/2021	archiver 9	1/9/2021	archiver 39
763eg41	Prothomalo	1/10/2021	10	Dhaka	10th	1/10/2021	archiver d	1/10/2021	archiver 40

Fig: Archived newspaper list

3.3 Update:

The interface should have the functionality where the archiver can update any uploaded data, including all the meta-data using the unique ID.

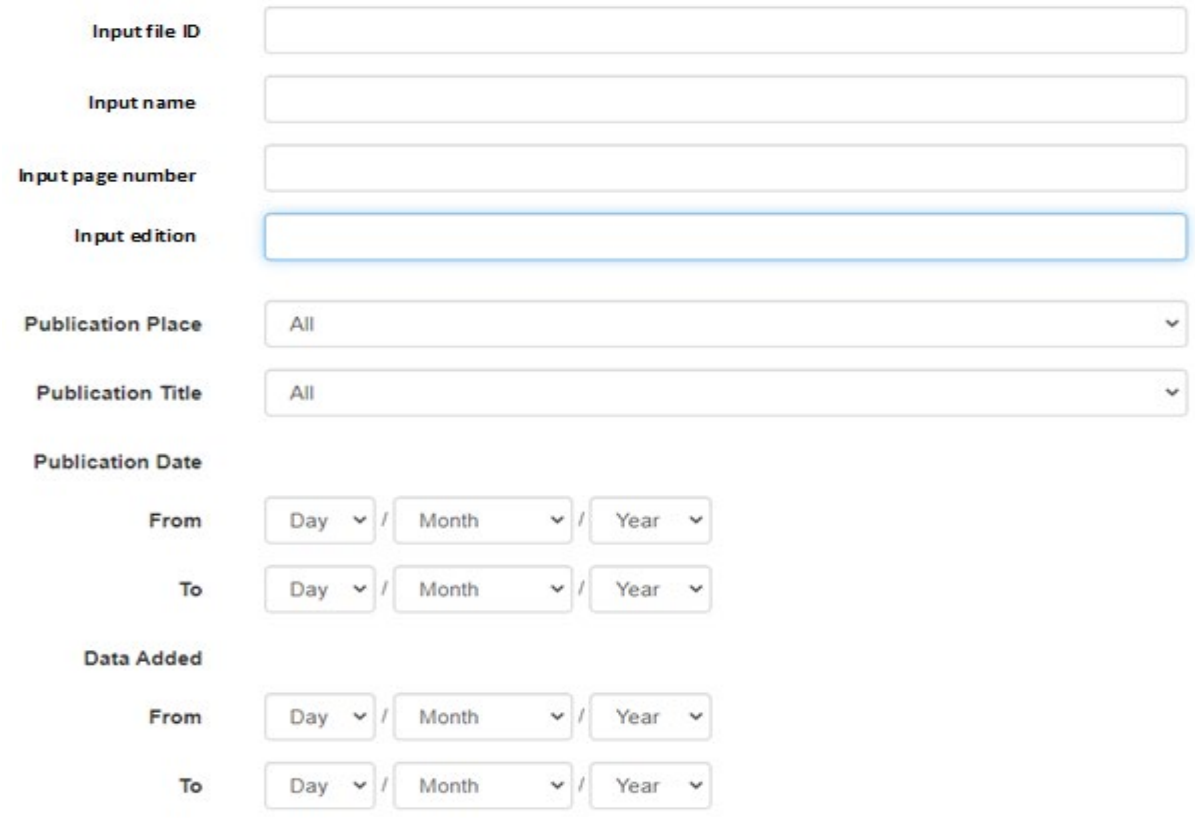

Fig: Archiver update interface

3.4 Delete:

Deleting any entry should require special permission, and also deleted items should be stored in the recycle bin for 2 months. Finally, clearing the recycle bin should also require special permission.

Fig: Archiver delete interface

Annexure-4

4 User Interface

The User interface should contain two main feature viewing and searching containing the following criteria explained below

4.1 View

Viewers should be able to find all the newspapers stored in the archive as a list. The list should contain all the data columns from categorization data, and the viewers should be able to sort and filter all the newspapers by the following criteria:

1. Name
2. Date
3. Page
4. Location
5. Edition

Name	Date	Page	Location	Edition
Prothomalo	6-Sep-21	1	Bangladesh	1 st
Prothomalo	6-Sep-21	2	Bangladesh	1 st
Prothomalo	6-Sep-21	3	Bangladesh	1 st
Prothomalo	6-Sep-21	4	Bangladesh	1 st
Prothomalo	6-Sep-21	5	Bangladesh	1 st
Prothomalo	6-Sep-21	6	Bangladesh	1 st
Prothomalo	6-Sep-21	7	Bangladesh	1 st
Prothomalo	6-Sep-21	8	Bangladesh	1 st
Prothomalo	6-Sep-21	9	Bangladesh	1 st
The Daily Star	6-Sep-21	1	Bangladesh	1 st
The Daily Star	6-Sep-21	2	Bangladesh	1 st
The Daily Star	6-Sep-21	3	Bangladesh	1 st
The Daily Star	6-Sep-21	4	Bangladesh	1 st

Fig: list of the newspaper

Any item from the list should be clickable, and clicking on any item should lead to the desired archived newspaper as shown below:

Fig: Newspaper

4.2 Search

The User interface should be built in such a way so the user should be able to give the name, date, page number, location, and edition of the newspaper as inputs in the different search box as criteria of the search in order to narrow down the search and a search button to initiate the search.

Similar to the below figure, the search box/form should contain all the search boxes. The user should be able to fill the boxes depending on their needs. Filling up all the boxes should not be mandatory. The search engine should take any input and show all the search results in the viewing list, and the user should be able to serf through their searches. The feature should help narrow down searches.

For OCR pdf, optical characters should have a search/find option for OCR PDF view.

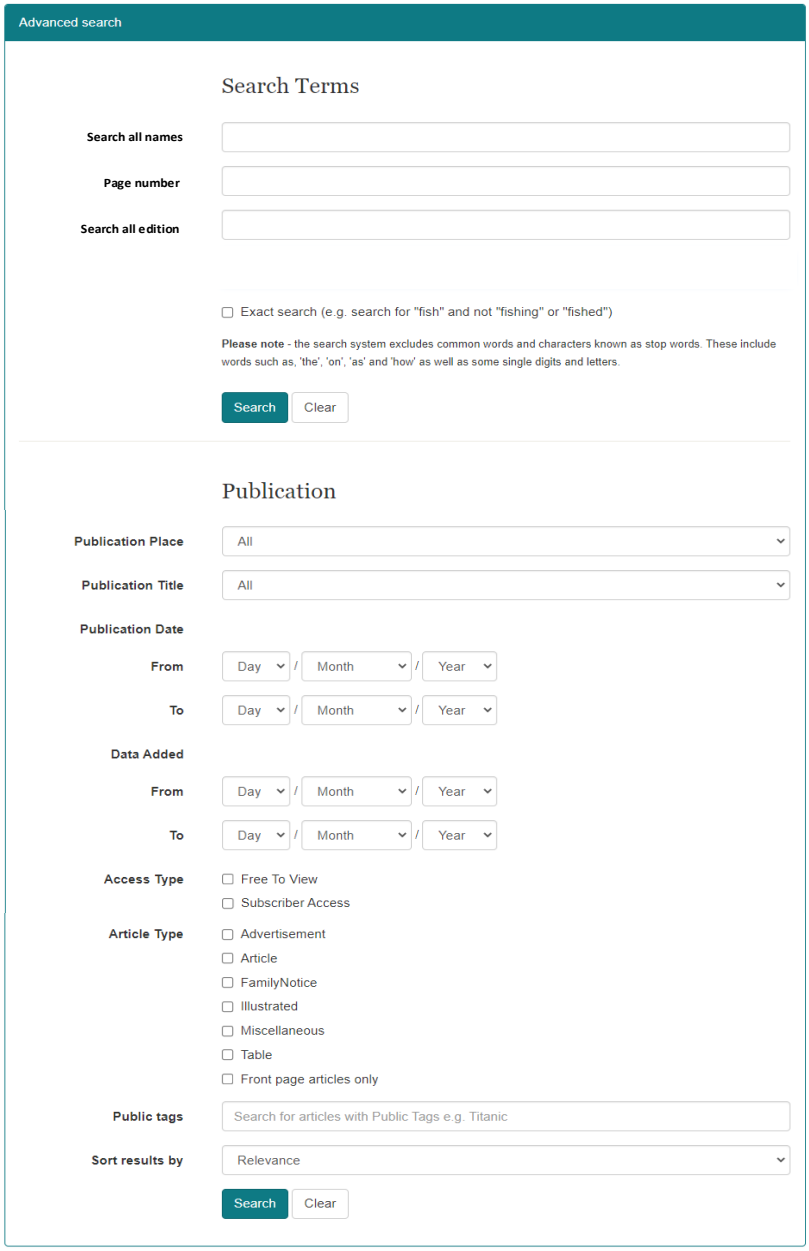

Fig: user interface for searching

Annexure-5

5 Systems architecture

5.1 Description

All the scanned data including all the additional data should be stored in the server, and for each new data in the server cloud backup should initiate and store the newly added data as a backup update. The User interface should be connected to the main server in order to fetch all the documented newspaper data plus some additional data for maximum user experience.

5.2 Diagram

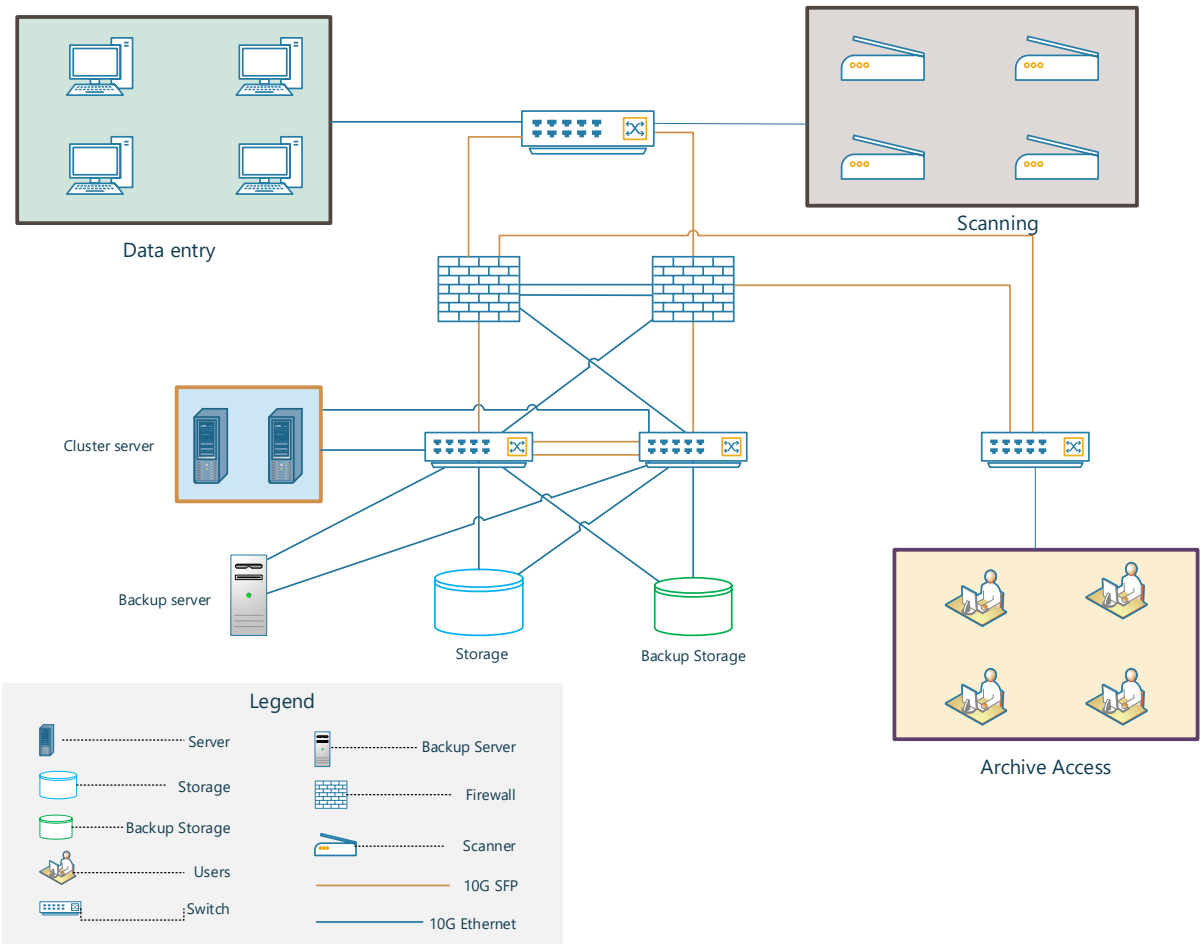

Fig: High-level Process Diagram

5.3 System requirement:

Item No	Descriptions	Required Specification
2 Units Servers for Application		
1.0	Processor Clock Speed	Minimum 2.1GHz
1.1	No of Processors	01 (One)
1.2	Processor Cores	Minimum 8 cores per processor
1.3	RAM	16 GB Memory with Expandable option
1.4	Internal HDD Capacity	1 TB
1.5	Ethernet Card	Min. Two port 10Gb SFP+ Network Card
Storage Requirement (Database)		
2.0	Storage Space	7.5 TB (Useable)
Backup Storage Requirement		
3.0	Storage Space	7.5 TB (Useable)
Database OS		
4.0	DB OS	PostgreSQL or MySQL on a high availability system
Backup Software		
5.0	Backup Software	Need a backup solution to protect human error while in preparing the file and finally a complete archive solution as a destination to keep those files forever.

5.4 System capacity and performance

System usage and allocation of resources should meet the below requirements

5.4.1 Latency for search and view:

To start with, while viewing all the newspapers in the archive list, updating/reloading the list while sorting and filtering should not take more than 1 second. On the other hand, any kind of search should not take more than 2 sec for the maximum number of users using the interface simultaneously. also

5.4.2 The number of users/viewers:

The system should be shaped in such a way that at least 200 users should be able to use the interface simultaneously without any lag, error, glitch, or server overload interruption. Among the 200 users, at least 50 of them should have not only viewing privileges but also upload, download, delete and overwrite privileges.

5.5 System security

Secure Software Development Lifecycle (SDLC) to be followed, and the whole system security should be maintained by ensuring the below holistic security standard practices:

5.5.1 Application Security Plan:

1. User login authentication should be implemented based on the requirements.
2. If an existing user of the system logs in using a new device, then system administrators should receive a notification with the IP address of the device.
3. Security features should include password protection.
4. The system should be completely secure and foolproof with the incorporation of industry-standard proven data encryption techniques and methodologies. Those encryption techniques should be audited in a timely manner to detect loopholes and updated with the latest patches in order to ensure that the mechanisms are fitted with the latest security features.
5. The system should block the user account for a parameter-driven length of time after a parameter-driven number of invalid login attempts.
6. The system shall provide definable password enforcement rules, including but not limited to:
    a. Password length
    b. Required alpha & numeric character
    c. Not the same as the previously used password
7. The access control security function should provide a facility for the system administrator to suspend an existing user's access rights for a specified period of time or indefinitely.
8. User role-based security should be implemented throughout the system. As a result, users should not be able to access all the links in the system.

5.5.2 UI Security Plan:

1. There should be a session time. After that time, it should be auto-logout, and the user may have to log in again. In the case of session logout, the full screen should be under a shadow, and a login panel should be shown. After login, the user should continue were left off. But in the case of user logout, the user logs in again and start from the beginning.
2. The system should be completely secure and full proof with the incorporation of industry-standard proven data encryption techniques and methodologies along with the implementation of Secure Socket Layer (SSL).

5.5.3 Application Scalability Scopes:

Both horizontal scaling (scale-out) and vertical scaling (scale-up) should be possible so that in different situations, the most logical steps can be taken. The system should provide an appropriate caching mechanism to handle very high-traffic scalability and must be optimized to deliver content to the end-user within a short response time.

5.5.4 Interoperability & Data Exchange Scopes:

1. The system should be designed for interoperability using industry-standard protocols.
2. All imported data should undergo data validation to ensure full integrity.
3. Data exchange within the system at different levels via the internet shall be encrypted on a need base.
4. The system should have the functionality to exchange data with other relevant databases in a most secure environment through a standardized data exchange protocol designed, developed and implemented.
5. The system should be able to maintain the log of such data imports being performed by the authorized users.

5.5.5 Backup requirement

1. Need a backup solution to protect human error while in preparing the file and finally a complete archive solution as a destination to keep those files forever.

5.5.5.1 Database Security Compliance:

1. Public Network Access to the Database Servers should be disabled
2. There should be two separate databases & application servers
3. Database backups should be maintained in publicly inaccessible locations
4. Regular Database Security Assessments should be performed
5. Physical Database Security for accessing the database should be ensured
6. A dedicated secure network for the application database should be ensured

5.5.5.2 Data Privacy Compliance:

1. All the data should be used in a test environment in a secured private environment when testing is required in the development phase
2. All the data should be used in a secure network and only used for the purposes of investigating the support case.

Annexure-6

6 Reference

1. https://pixelcalculator.com/en
2. https://veridiansoftware.com/knowledge-base/newspaper-scanning/
3. https://www.synopsys.com/blogs/software-security/secure-sdlc/#:~:text=A%20software%20development%20life%20cycle%20(SDLC)%20is%20a%20framework%20for,speed%20and%20frequency%20of%20deployment.
4. https://www.opticallimits.com/jpeg2000-vs-jpeg-vs-tiff
5. https://www.newspaperclub.com/create/design-guides/image-resolution