# 6. Content Integration

## 6.1. Need for Integration

In the context of State Portal "Content" consists of all the data proposed for being published. Published content would be referred as "Information". Content could be authored on State Portal or sourced from government sources or private parties or through live feeds. Content would be structured such as records for database, XML files or unstructured such as images, documents (all in formats PDF, DOC. Postscript etc.), HTML files, audio and video.

Content authoring refers to creating content objects on State Portal. Content authoring related functionality and implementation is explained in "content management framework" section.

Each State Portal would have thousands to millions of web pages and documents. These documents would be related to government matters and contain wealth of information required by citizens, businesses and overseas people. There will be hundreds of portals from government departments, organizations, universities, district collector, blocks etc. In a scenario where every portal or website is governed in its own way, citizens will find it extremely difficult to access or discover the information they are looking for. They are likely to get frustrated in hopping from portal to portal. In order to provide citizens an easy access to information an integrated approach towards making content easily discoverable, diminishing semantic inconsistencies and avoiding content duplication is a must. Integrated approach will enable easy access to information and make all content published on government websites, especially State Portals, easily discoverable.

In order to provide integrated information, State portal need to automate the processes of

a. Content sourcing (feeds and documents)
b. Content aggregation or delivery
c. Content discovery

Following sub-sections explains each of these scenarios.

## 6.2. Content Sourcing

### 6.2.1. Content Sourced in Paper Form

Plenty of government content will be generally published in paper form rather than in electronic form. State Portal would facilitate preparation of electronic copy (document digitization) of this content, which includes providing scanned images of paper documents. However ownership of document digitization lies with state departments and organizations. In this scenario following guidelines should be followed:

a. In order to gather information Content authoring team should adhere to defined procedures to reach out various government and third party information.
b. By means of metadata and compliance processes, reference to original source should be maintained.
c. Defined content publishing workflow should be followed for publishing electronic copy of content on State Portals so that required metadata can be associated with the content.
d. Audit and compliance processes should be defined to validate timeliness of content, correctness and completeness of metadata.

### 6.2.2. Content Sourced From Other Government Websites

Information delivered through State Portal would also be sourced from many government websites. New content object would be created by means of summarizing the content of other government website. In this type of scenario following guidelines should be followed:

a. Defined publishing process should include extraction of metadata from the sourced website. Referenced URL should always be maintained as part of content.
b. Audit and compliance processes should be defined to validate timeliness of content, completeness and correctness of metadata.
c. Compliance processes should validate content correctness and consistency of metadata between original website and State Portal.
d. Content authoring team should adhere to defined procedures to reach out various sources of citizen related information.
e. Crawling process should be defined to discover content useful for citizens and publishing the same on State Portal.
f. Content archival policy should validate consistency between original content source and State Portal.
g. Compliance processes should include periodic checking for the existence of referred URLs, and notify content administrator about missing references.

### 6.2.3. Content Sourced From Third Parties

Some of the information provided on State Portal would also be sourced from private or third parties. In this type of scenario following guidelines should be followed:

a. Content must be sourced only in electronic form.
b. Contract must be created for all third party content providers.
c. Appropriate additional quality checks should be incorporated as part of content publishing process to ensure quality of content.
d. Verification with respect to adherence to defined contracts should be performed.
e. Appropriate content existence and archival policy should be defined

### 6.2.4. Live Feeds

State Portal would be sourcing content in the forms of feeds as well as providing feeds. Functionality and implementation guidelines for live feeds are provided in “Content management framework” section.

### 6.2.5. Providing Links to External Sources

State Portal would work as a consolidated repository of all state government information provided from all state level departments and organizations. Original source of most of the information provided on State Portal will be departments or organizations. However this does not mean State Portal will have all the information provided on websites or portals of all state level departments and organizations.

In some cases content objects created on State Portal would refer other government websites by means of just providing a link to other government websites, so that for further details users can be redirected (if required). In this scenario no content would be duplicated or summarized. In other words URL will form part of the content object. In this scenario following guidelines should be followed:

a. Compliance processes should include periodic checking of the existence of referred URLs.
b. If referred URLs are found to be broken then content administrator should be notified about missing references.

c. This type of content sourcing should only be used for providing further reference rather than providing content from other websites.

## 6.3. Content Delivery

### 6.3.1. Content Delivery within State Portal

State Portals would have large number of content objects and documents. It must manage complete life cycle of all content objects, for this purpose, it must use a content management system. Content management systems include content aggregation/deployment tools, which make use of content objects from content management system's content repository.

### 6.3.2. Content Delivery using multiple State Portal's content repositories

State Portal would have authoring tools to aggregate content from its content repository and publishing it on State Portal. At present due to absence of open standards content aggregation tools will find it difficult, to make use of content from multiple content repositories, which are not part of same content management system. Therefore, content delivery using multiple State Portals' content repositories would not be mandatory.

## 6.4. Content Discovery or Searching

### 6.4.1. Discovering Information within State Portal

State Portal should provide metadata and 'full text search' based search functionality. Search functionality should be exposed as a web-service so that other State Portals can provide integrated search functionality. For providing search functionality State Portal should comply within defined processes for defining metadata, managing metadata schema changes and master data changes.

### 6.4.2. Discovering Information from all State Portals

There would be situations where users would not be aware of which State Portal or government website would serve their needs. There would be some scenarios where in users like to do some sort of comparison and analysis based on the similar information retrieved from multiple State Portals. Therefore State Portals should provide the functionality where in users would be able to search across all State Portals. To provide this functionality, metadata should be consolidated from all State Portals on periodic basis.

## 6.5. Content Integration Approach

Content integration will require number of processes during the entire life cycle of content. This section describes the State Portal's approach for content integration.

### 6.5.1. Standardize Content Taxonomy

Content taxonomy should be standardized. State Portal would comply with defined standard. Compliance and governance processes should check adherence to defined standards. Content publishing workflow should enforce taxonomy rules as part of publishing process.

### 6.5.2. Standardize Metadata Schema

Metadata schema should be standardized. State Portal would comply with defined standard. Compliance and governance processes should check adherence to defined standards. Content publishing workflow should enforce definition of all metadata attributes as part of publishing process.

### 6.5.3. Standardize Master Data

Master data should be standardized. State Portal would comply with defined standard. Compliance and governance processes should check adherence to defined standards. Content publishing workflow should enforce usages right master data values as part of publishing process.

### 6.5.4. Standardize Content Publishing Workflow

Content publishing workflow should be standardized web enabled automated process. It is expected that different State Portals may use different products for content management. Definition of metadata should be integrated into publishing workflow. Wherever possible “free flow text” should not be allowed as metadata values.

### 6.5.5. Standardize Content Management System

All State Portals should use a content management system. Content management system should provide defined set of standard functionality.

### 6.5.6. Unique Content Identifier

All content should be assigned a unique ID. Content repository should have web based access. All published content should be accessible using a URI.

### 6.5.7. Consolidated Metadata Repository

A consolidated metadata repository should be established, which will store metadata of all content published on all State Portals using an automated process. State Portal should send metadata of newly added content and metadata changes of existing content, periodically to the consolidated metadata repository. In order to support interoperability a web-service should be defined. This web-service will receive metadata from State Portals, normalize it and save to the consolidated metadata repository.

### 6.5.8. Metadata Normalization

Metadata should be normalized as part of metadata consolidation process. Normalized process should map state specific variation of metadata values to standard terms. This will enhance quality of search results for consolidated content search functionality.

### 6.5.9. Metadata Schema Changes

Metadata schema changes should be managed through defined change management process as part of State Portal governance. Processes should ensure that changes are applied on all State Portals and the consolidated metadata repository.

### 6.5.10. Master Data Changes

Master data changes should be managed through defined change management process as part of State Portal governance. Process should ensure that changes are applied on all State Portals and consolidated metadata repositories.

### 6.5.11. Metadata Updates

When new content gets published, its metadata should be propagated to consolidated metadata repository. Similarly metadata may get changed after publishing of content. In such case also changes should be propagated to consolidated metadata repository. Propagation of metadata should be managed using an automated process. Availability of this process should be monitored as part of compliance processes. It should guarantee propagation of all metadata updates to consolidated metadata repository.

### 6.5.12. Standardize Web Based Interface to Content Repository

Following guidelines should be followed:

a. Web based interface should be provided to access and browse content repository.
b. Following functionalities should be provided:-
    i. Querying content objects
    ii. Browsing content objects
    iii. Querying metadata based on various criteria's such as state, modified since etc.
    iv. Browsing metadata
c. Exporting metadata in XML and other open standard formats
d. Defining content exit policy
e. Defining content archival policy

## 6.6. State Portal Search Service

State Portal would have very high volume of information which includes documents, HTML pages, images, audio files, video files etc. State Portal users would also be from diverse backgrounds. Therefore, it is must that portal provides a highly user friendly search functionality.

Following would be the key search functionality

a. Citizen and portal users should be able to search within a State Portal
b. Provide unified user interface to perform above search, through 'full text search' capability
c. Provide metadata based search capability
d. Provide explorer type of user interface to explore content

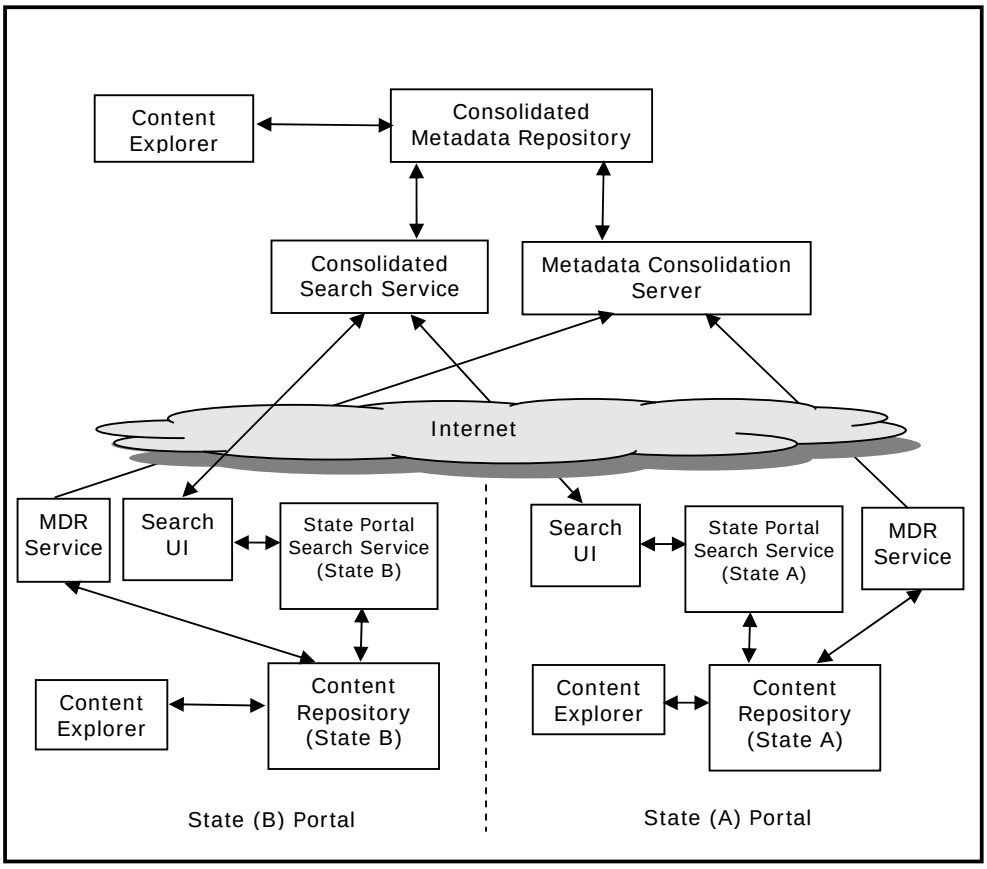


**Figure 9. Consolidated Search Service**

### 6.6.1. Search User Interface

In order to provide easy access to information, it is essential that users will be able to discover the desired government content using a single window interface and content of all State Portals should be made discoverable. Users should be able to explore the content from a single state as well as multiple states in seamless manner. Further all State Portals should provide similar user experience.

State Portal would provide following user interface for search functionality:

a. Metadata based search and full text search
b. Searching within State Portal
c. Uniform user interface to perform above types of search

d. Content explorer

### 6.6.2. State Portal Entities

#### 6.6.2.1. State Content Repository

State Portal is expected to have volume of information. Content belonging to State Portal would be stored in content repository of State Portal. State Portals would use a “Content management system” to store and manage content.

State content repository would have defined standard functionality. Functionality expected from content management system is detailed in “Content management framework” section.

#### 6.6.2.2. Content Explorer

Content explorer is a web application for browsing the content of State Portal. It would also be available to browse the content of all State Portals using “consolidated metadata repository”. This application will be similar to ‘explorer’ application available on Windows operating system or browsing application provided by analyst site like forrester.com.

Content explorer would have a defined standard user interface so that citizens will have same experience on all State Portals.

#### 6.6.2.3. Search User Interface (UI)

This represents the web based interface for search functionality provided by State Portals. This would be implemented using technologies like JSP, Servlet, PHP, ASP.Net etc.

Search user interface would confirm to a defined standard so as to provide unified and standard user experience to citizens from all State Portals.

#### 6.6.2.4. State Portal Search Service

State Portal would expose search functionality as a service so that other State Portals or government portals can invoke it to search the content available on the State Portal.

Service would have a well defined standard interface so that citizen would have unified and standard user experience on all State Portals.

### 6.6.3. Shared Entities

#### 6.6.3.1. Consolidated Metadata Repository

Consolidated metadata repository would store and manage metadata of all content published on all State Portals. It will store metadata in normalized form. Normalization will enable high quality search results and provide semantic integration of content from State Portals.

Consolidated metadata repository would have a well defined standard metadata schema and normalized values for metadata based on well defined standard master data.

#### 6.6.3.2. Consolidated Search Service

Consolidated search service would provide metadata based search functionality to search all State Portals. It will be implemented as a service and deployed in a shared environment so that all State Portals can easily access it. It would use “consolidated metadata repository” as its content store.

Consolidated search service would have a well defined WSDL compliant standard interface.

## 6.7. Metadata Replication Service (MDR Service)

MDR service would be an integral part of State Portal. It is a must for:

- Realizing metadata based content integration
- Consolidating metadata from all State Portals
- Integrating State Portals with National portal and
- Providing metadata based content discovery functionality

It would implement a well defined interface providing standard set of functionality as mentioned below. Its implementation may wary from State Portal to State Portal depending on the specific content management and technologies used for implementing State Portal.

MDR service must provide following functionality

a. It should implement a well defined WSDL compliant standard interface.
b. It should be implemented as a web service
c. It should be deployed on a highly available infrastructure.
d. Indicate it's availability and normal functioning.
e. Provide means to determine
    i. Published content since given time
    ii. Published content, whose metadata is modified since given time
    iii. Published content, which is exited since given time
f. Provide means to query metadata from State Portal's content repository
g. Retrieve metadata of given published content from the content repository of State Portal, satisfying each of following criteria
    i. Published content since given time
    ii. Published content, whose metadata is modified since given time
    iii. Published content, which is exited since given time
h. Normalized metadata based on defined standard master data.
i. Send incremental metadata changes to 'consolidated metadata repository' through 'metadata consolidation server'.
j. Maintain time upto which metadata and metadata changes were propagated to "consolidated metadata repository".
k. Ensure that propagation of metadata and metadata changes even when
    i. MDR service temporarily goes down due to any reason, including
        1. Internet connection failure
        2. Hardware failure
        3. Operating system failure
        4. Web server
        5. Application server failure
        6. Content repository failure
        7. Web service software failure etc.
    ii. 'Metadata consolidation server' goes down temporarily or fails to invoke "MDR service"
l. Ensure that all metadata and metadata changes are propagated only once.

MDR service should be implemented such that code specific to content repository and content management system is well encapsulated so that migration from one content management system to another content management system should have

- No impact on "metadata consolidation server"
- Minimal impact on "metadata replication service"

### 6.7.1. Development and Deployment Guidelines

MDR service may be developed as a web service or a SSDG service or both. In case of web service, it would be deployed on a web service compliant service communication infrastructure. In case of SSDG service, it would be deployed on a SSDG compliant service communication infrastructure.

## 6.8. Metadata Consolidation Server (MDC Server)

Metadata consolidation server would be a UNIX daemon or NT service type of entity, which would be running all the time. It would provide well defined standard interface to communicate with “metadata replication services”, which are integral part of State Portal.

MDC server would provide following functionality:

a. Perform conformance verification of metadata and metadata changes received from “MDR services” with well defined standard metadata schema
b. Perform conformance verification of attribute values of metadata and metadata changes received from “MDR services” with well defined standard master data
c. Check availability and normal working of registered MDR services at defined periodicity
d. Retrieve metadata and metadata changes from all state portals at defined periodicity using registered MDR services
e. Maintain list of MDR services or State Portals from which metadata to be consolidated
f. Maintain time upto which metadata is received from each MDR service
g. Provide alerts for various errors such as
    i. Non conformance to defined standard of metadata schema
    ii. Non conformance to defined standard of master data
    iii. Failing to normalize received metadata
    iv. Failing to save received metadata to “consolidated metadata repository”
    v. Failing to invoke MDR services
h. Provide means to take one or more of following actions on detection of above errors
    i. Log errors to defined log file or database
    ii. Display message on system console or event log
    iii. Send email message to defined email IDs (such as web master of State Portals)
    iv. Send escalation email message if error condition persists beyond defined amount of time.