ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Short Communication

Open Access

Importance of ICC in Modern Content Extraction and Bulk Upload

Ravi Kiran Kanneganti

Manager, Capgemini America Inc, USA

***Corresponding author**
Ravi Kiran Kanneganti, Manager, Capgemini America Inc, USA.

Received: February 20, 2022; **Accepted:** February 25, 2022; **Published:** March 05, 2022

IBM Content Collector for Email (part of the IBM Content Collector family) is an automated, policy-driven solution primarily used for archiving emails to a central Enterprise Content Management (ECM) repository.

The product's ability to extract data from emails is a core part of its archiving process, which enriches the email with metadata to enable efficient search, compliance, and lifecycle management.

How Data Extraction Works for Emails

The extraction process in IBM Content Collector for Email is performed by the EC Extract Metadata task within a defined Task Route. Its focus is on harvesting the structured and unstructured data from the email to facilitate later retrieval and governance.

A. Metadata Extraction

The most critical part of data extraction is the automatic harvesting of metadata from the email header and properties. This includes:

- **Email Fields:** Sender (**From** Address/Display), Recipients (**To, CC, BCC** addresses/displays), **Subject** line (including a "Conversation Topic" cleaned of reply/forward prefixes), **Date Sent/Received**, and **Attachment Count**.
- **System Properties:** Internal mail server IDs, message size, folder path, and flags (e.g., whether the message is encrypted or signed).
- **User-Defined Metadata:** Administrators can configure the system to extract data from **additional fields** beyond the defaults, enabling custom indexing based on specific business needs.
- **B. Content and Text Extraction**
- The system extracts and processes the actual content of the email and its attachments for full-text search capabilities:
- **Email Body:** The complete text of the email message is extracted.
  Attachment Content: A Text Extraction Connector (which often uses Oracle Outside in Technology filters) converts the binary data of attachments (like PDFs, Word documents, and spreadsheets) into a plain-text representation. This makes the **attachment content searchable** in the central repository.

Primary Goals of Email Data Extraction

The extracted data is not just for viewing; it serves essential **information governance** and **IT efficiency** goals:

Goal	Description
Search and eDiscovery	The rich metadata and full-text extraction allow users and legal teams to perform eDiscovery searches based on a combination of fields (e.g., "All emails from Smith about 'project X' sent in 2024").
Retention and Compliance	Extracted dates, sender, and content keywords are used to trigger policy-driven archiving and apply specific retention schedules to individual messages, ensuring regulatory compliance.
Storage Management	By knowing which emails are duplicates (Deduplication) or which can have their body/attachments removed (Stubbing), the system uses the extracted data to manage and reduce the overall storage footprint on email servers.
Workflow Integration	Extracted data can be used to route the archived email or to initiate a business process in a connected system (like an ECM or BPM platform).

IBM Content Collector extracts metadata from **XML files** primarily using the **File System Connector (FSC)** and the **FSC Associate Metadata** task within a task route. This is typically used when you have content files (like PDFs or images) in your file system and separate XML files containing descriptive metadata about those content files.

Key Components for XML Metadata Extraction

The process is centered on defining a mapping between XML elements and internal Content Collector metadata properties.

The FSC Associate Metadata Task

The FSC Associate Metadata task is the core component for this function. It works by:

- Matching: It associates a content file (e.g., invoice.pdf) with its corresponding metadata file (e.g., invoice.xml) based on a predefined rule (often a matching name with a different extension).
- Extraction: It reads the associated XML file.
- Mapping: It extracts specific data points from the XML using **XPath expressions

IBM Content Collector extracts metadata from PDFs primarily as part of its File System Archiving process or when the PDF is an email attachment. It achieves this through two main mechanisms within the task route: internal text extraction for content and associating external metadata files.

- **Extracting Text Content and Basic Properties**
  For any document, including a PDF, Content Collector's internal processes handle the file itself:
- **Text Extraction:** Content Collector uses an internal component, often powered by **Oracle Outside in Technology** filters (via the **Text Extraction Connector**), to convert the binary data of the PDF into a plain-text representation. This text is stored alongside the document in the repository (like FileNet P8 or Content Manager) and is crucial for full-text search and eDiscovery.
- **System Metadata:** When a PDF is collected (e.g., from a monitored file share via the **File System Connector**), the system automatically captures basic file-level system metadata:
  - o File Name and File Path.
  - o Date Created and Date Modified.
  - o File Size and Content Type.

Associating Custom or Business Metadata

While Content Collector can extract the text, its main mechanism for enriching a PDF with business-specific metadata (like "Customer Name" or "Invoice Number") is through its file system collection tasks:

A. FSC Associate Metadata Task

If a PDF file is accompanied by a separate metadata file (typically in XML or CSV format) that contains its key business data, the FSC Associate Metadata task is used.

- Process:

1. The File System Collector (FSC Collector) collects the PDF (e.g., invoice_12345.pdf).
2. The FSC Associate Metadata task locates the corresponding metadata file (e.g., invoice_12345.xml).
3. It reads the data from the XML or CSV file (e.g., <customer>ABC Corp</customer>) and associates that data with the PDF document as custom metadata properties in the ECM repository.

B. FSC Metadata File Collector

Alternatively, an administrator can use the FSC Metadata File Collector to drive the process entirely from the metadata file.

- Process:

1. The Collector is configured to monitor for XML or CSV metadata files.
2. It reads a metadata file that contains the path and name of the corresponding PDF document.
3. It uses the information in the metadata file to collect and archive the PDF, and it applies all the data from the metadata file as custom properties to the archived PDF.

This flexible approach allows organizations to integrate with upstream systems that generate business-critical data separately from the content file.