



INTEGRATING CONTENT INTELLIGENCE USING INFOSYS COGNITIVE TAGGING PLATFORM

1 Introduction

Post covid-19, the world of interactions demands 100% digitization. There is more hunger in the world for process automation, triggered by the new normal. During process automation, business decisions are primarily taken based on the information provided by the user and are validated automatically against digitized proofs.

Let us look at a sample scenario of KYC. In a KYC application, a customer is uploading his/her passport. But by mistake, the customer selects the wrong document while uploading. The application should be intelligent enough to identify the uploaded document is not a passport and alert the customer before accepting the uploaded document. Further, let's assume the customer uploaded the correct document, the system should be intelligent enough to extract all important metadata or information from the uploaded passport and pre-fill the metadata values or information on the user interface. The customer will just verify the information and submit the information for further processing. Advanced intelligence can be processing the extracted metadata and conducting a few checks automatically, such as, passport expiry date should not be a past date, etc.

In all above scenarios, content intelligence is required, to take decisions or to automate processing. In the first scenario, the

system took the decision that the uploaded document is not passport and alerted customer accordingly. In the second scenario, the system pre-fills all metadata to automate the key-in activity.

Surely, customers will be delighted to see such "Intelligent user experience" while interacting with an application. To implement such intelligent applications, a content intelligence layer is needed. The content intelligence layer will derive the intelligence from the content, which then can be used to implement intelligent user experience.

"Infosys Cognitive Tagging Platform" or ICTP, a Natural Language Processing (NLP) based solution, can derive content intelligence from the documents or unstructured content and then can be used further, as described in the above scenarios. The flexible content intelligence capabilities provided by ICTP can be seamlessly glued to any application – web or mobile experience. ICTP can be accessed through API to get the desired intelligence from the content and also it has its orchestration layer (or pipeline) to pull and process content on its own.

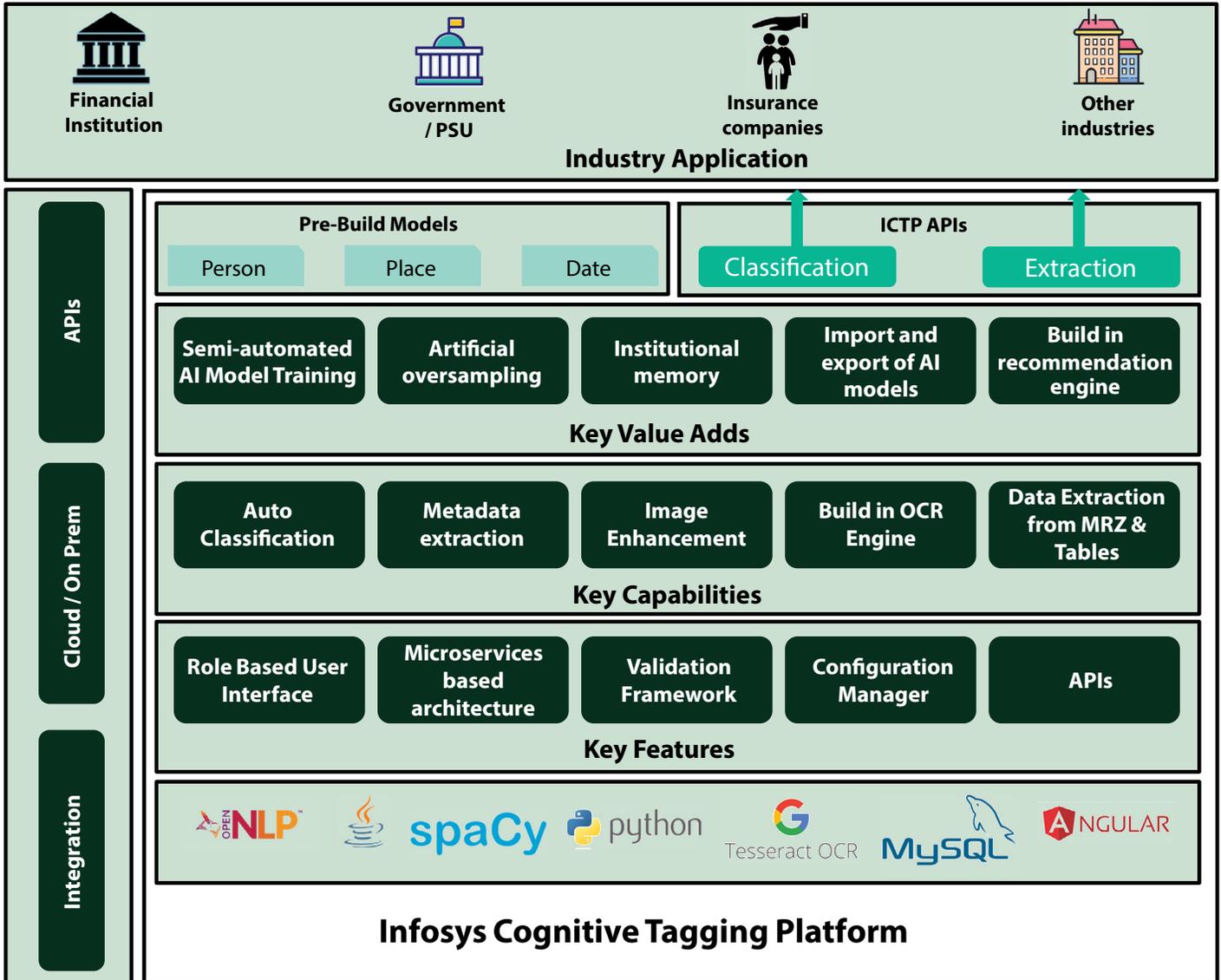
This white paper provides an overview of ICTP and how it can create a content intelligence layer to add "WOW" factor in simple applications.



2 What is Infosys Cognitive Tagging Platform (ICTP)

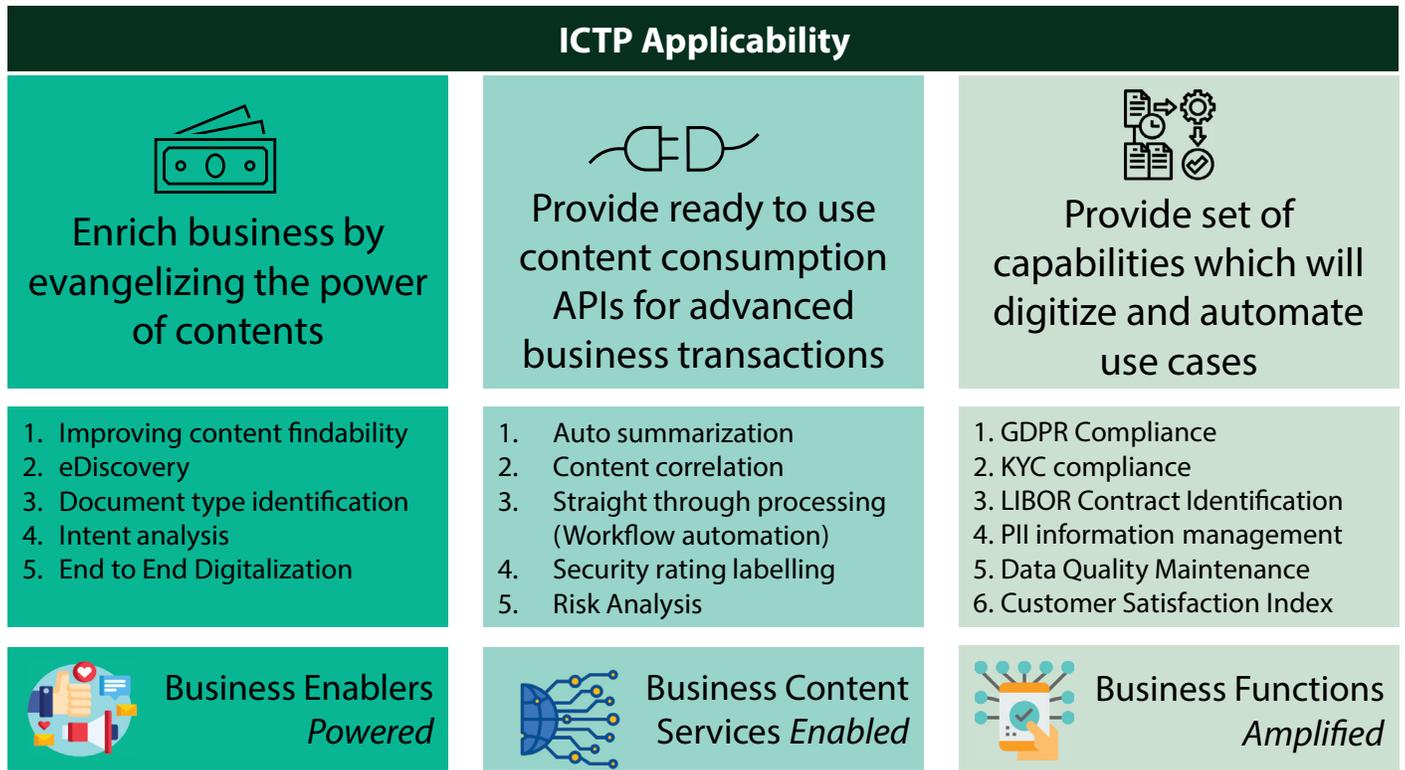
Infosys Cognitive Tagging Platform (ICTP) is a Natural Language Processing (NLP) based solution built using open-source tools which can derive content intelligence from unstructured content through classification and entity extraction. The platform provides user friendly interfaces, scalable platform in a microservices model and an API layer to provide tailored experience.

The following diagrams show components of ICTP.

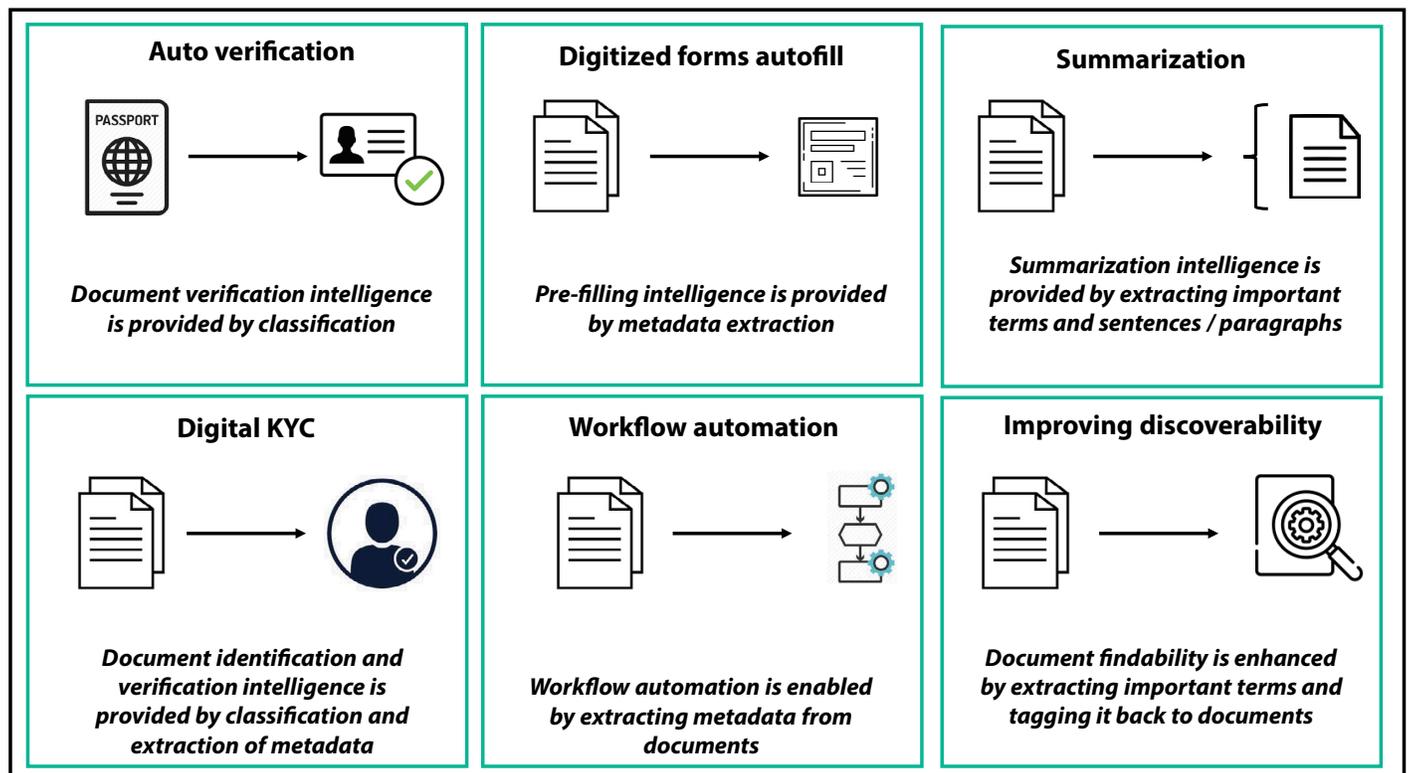


2.1 How ICTP can add business values

The applicability of ICTP can be summarized using the following diagram



Main deliverable of ICTP is to derive the content intelligence from the content. The derived content intelligence can then be used as follows.



2.1.1 Auto verification

ICTP can verify the type of document. By verification, the system can ensure that it is accepting correct documents

2.1.2 Digitized Forms autofill

ICTP can extract the metadata from content. The extracted metadata can be pre-filled on digital forms. User can quickly verify the attribute values and submit the digital form

2.1.3 Document summarization

ICTP can create summaries of large documents, contracts, or agreements, so that the user can go through the summary and understand it.

2.1.4 Digital KYC

ICTP can automate digital KYC by validating documents and verifying its authenticity by extracting metadata and connecting with external system.

2.1.5 Workflow automation

ICTP can automate the workflow by extracting desired attribute values and providing it to workflow engine. Workflow engine can use the extracted values to automate the workflow processing.

2.1.6 Improving document discoverability

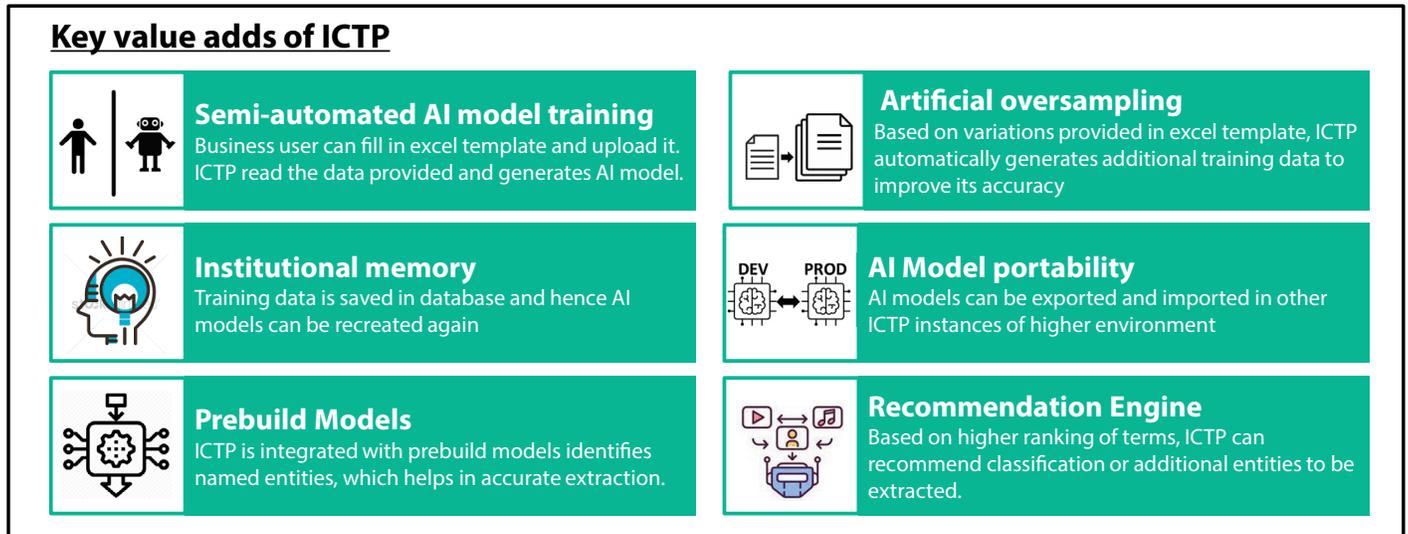
ICTP can extract addition terms from the document and additionally tag it to the document. This improves the chances of documents getting returned in an appropriate query.

Some of the use cases above are discussed in detail in a later section.



2.2 Key Value adds of ICTP

Key value adds of ICTP, as capabilities or features are as follows



2.2.1 Semi-automated AI model training

In any AI implementation, AI training takes significant time. To reduce the training time, ICTP uses a predefined excel template.

For training classification AI model, user needs to identify business terms which can be used to classify the document. Once listed in the template, the template can be uploaded to ICTP. ICTP then reads the information from the template, stores it into the database and then builds classification AI model from the information

For training information extraction, the process of training AI model is called annotation. Annotation is done using a user interface where user needs to identify the term that need to extract and tag it with an entity name. In ICTP, annotation can be done using a predefined excel template. Users need to add entity name and add a value of the entity and then add file path which is used as reference. It is mandatory that the reference file should contain the entity values, otherwise ICTP rejects the uploaded data during model building.

Filled in excel template and reference files can then be uploaded to ICTP. ICTP read data from excel file and convert the reference file in text format. Using the data in excel file and reference files, ICTP creates annotation file and builds AI model for information extraction.

With semi-automated AI training, the whole cycle of training is reduced to weeks from months.

2.2.2 Artificial oversampling

To increase the accuracy of extraction, it is necessary to provide as many variations in entity values as possible. More the variation in entity values more will be the accuracy. ICTP uses artificial oversampling technique to automatically create more variations in entity values. For every entity and its variation in value, ICTP generates one set of annotation file. If there are more than one variation, ICTP combines it with other entity and generates more annotation files. For example, if there is one entity "account_number"

with two variations, one with 10-digit and one with 8-digit. ICTP internally creates one set of annotation files with 10 digit account number and another set of annotation files with 8 digit account number. Both these sets of files are then used to train NER model. So, if there are two entities and both entities have two variations, then a total of four annotation files will be generated and will be used for training, which in term translate to better accuracy.

2.2.3 Import and Export of AI Models

In ICTP, AI models can be exported from lower environments and imported into higher environments and vice versa. Using this feature, AI models built in lower environments can be seamlessly deployed into production environments.

The feature of export and import of AI models makes ICTP truly agile. AI models can be quickly re-trained for new or unseen documents, tested and released in production.

2.2.4 Build in Recommendation Engine

ICTP has a built-in recommendation engine which provides recommendations for documents which are not classified using the existing AI model. Recommendation engine suggests important business terms that an unclassified document has. Users can select appropriate terms and assign a category and re-train the model. Once retrained, using updated AI model, ICTP can correctly classify the document, which was previously unclassified.

2.2.5 Institutional memory

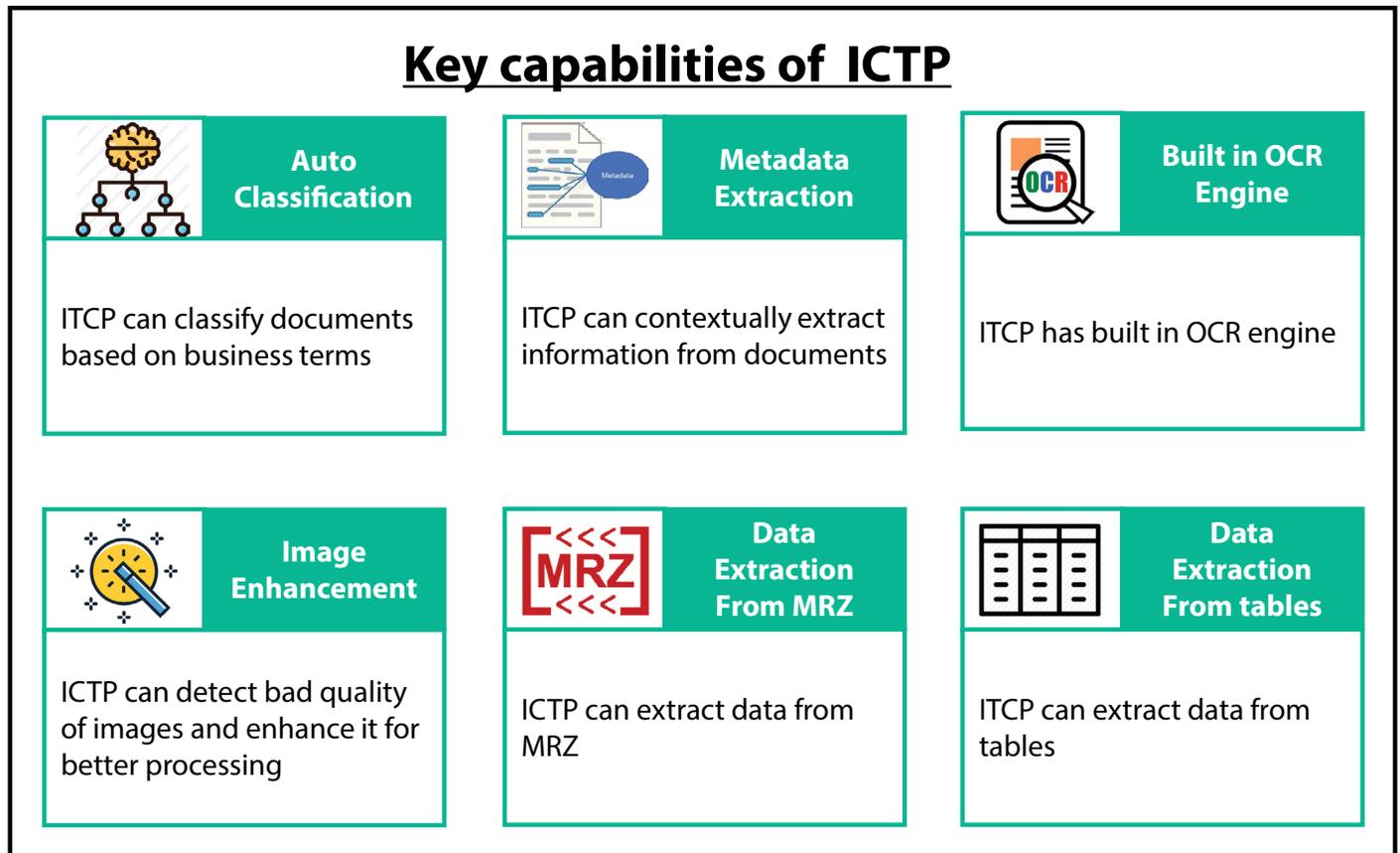
Whatever the data and files uploaded in ICTP, is all saved into the backend database. Hence, ICTP can re-create AI models at any time.

2.2.6 Pre-built Models

ICTP is integrated with prebuilt models which can identify named entities to increase accuracy.

2.3 Key Capabilities

ICTP offers end to end, “content intelligence deriving” solution by use of following key capabilities



2.3.1 Content Classification

Based on the context of the document, ICTP can classify content. To enable ICTP to classify content, prior training is needed. For classification, ICTP uses Apache OpenNLP as its NLP engine.

2.3.2 Metadata Extraction

Using Named Entity Recognition (NER), ICTP can extract contextual metadata from the content. These extracted metadata values can be used as per requirement.

2.3.3 Image Enhancement

ICTP uses python libraries to detect the poor quality of images and try to enhance it. With the inbuilt image enhancement module, ICTP delivers better classification and extraction results.

2.3.4 Built in OCR

ICTP uses open source tesseract OCR engine to get OCR the digital image and get the text from the image.

2.3.5 Integration with Pre-built models

There are many proven NLP models which can identify named entities such as Person, Place, Date etc. There are multiple ways, how output of these pre-built modes can be used. As of now, ICTP leverages these pre-built models to enhance the accuracy of its classification and extraction.

2.3.6 Data extraction from machine readable zones (MRZ)

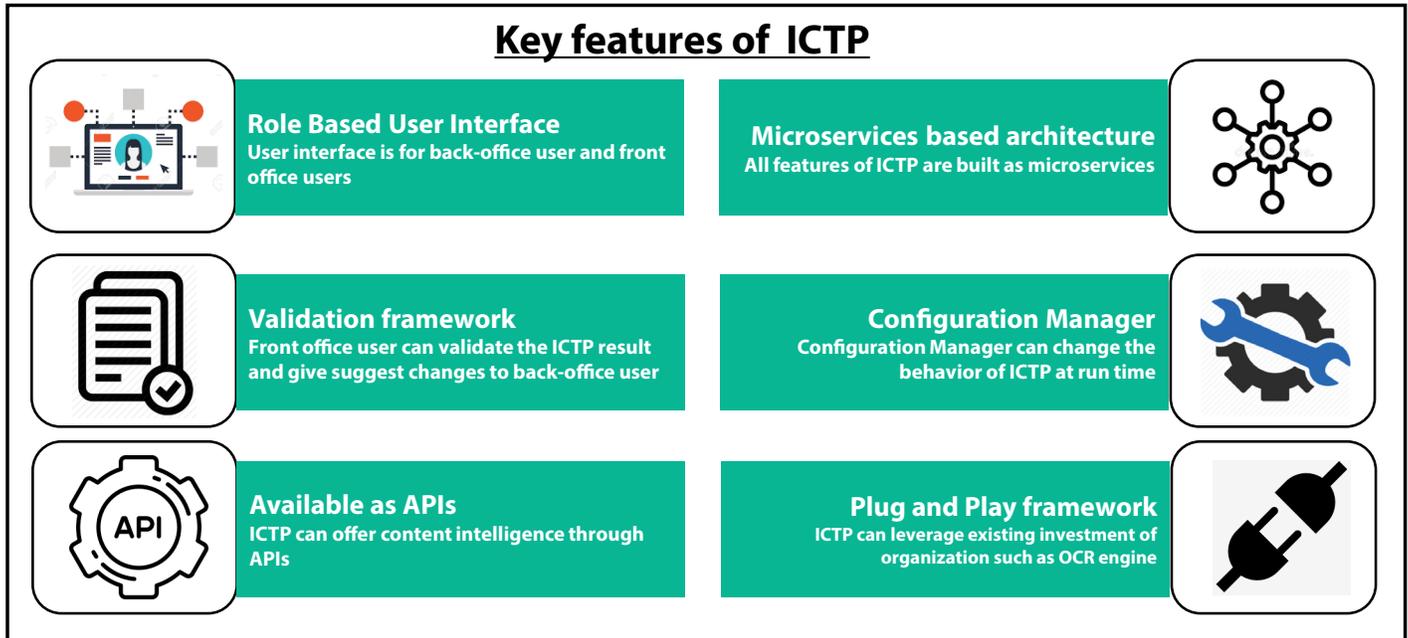
To extract metadata from machine readable zone (MRZ), ICTP uses python libraries, which automatically detects MRZ and extracts data from it.

2.3.7 Data extraction from tables

ICTP uses python libraries to detect the presence of table on document and then extract the information from the table as needed.

2.4 Key Platform Features

The following are key features of ICTP



2.4.1 User Interface (Role Based UI)

ICTP has a role-based, intuitive user interface. ICTP UI abstract all AI complexities and user just must do few clicks to make ICTP run.

User interfaces are designed based on generic role such as back office and front office. Back office is assumed to be IT users which mainly create AI models by uploading excel sheet and approve any changes in existing AI models. Whereas front office are business users who run AI model on unseen or new documents. When ICTP is run on new or unseen documents, front office user can see real time dashboard of what happening about the processing. On validation screen, front office user can validate the result of execution and suggest any business term addition in current AI model.

2.4.2 Microservice based architecture

Every feature of ICTP is implemented using one or multiple microservice(s). ICTP can be deployed in microservice architecture to scale the required throughput.

2.4.3 Validation Framework

ICTP has validation framework, where front office user can validate result of a batch execution. Using validation framework, front office user can reclassify a document and suggested new business terms for classifying future such document. Front office users can also select the business terms suggested by the recommendation engine. New suggested terms are sent to back-office users. Back-office users can approve these new business terms and re-train the AI model again. The validation framework is also available for information extraction scenarios.

When to kick in validation framework is configurable. When the AI model is matured and performing as per expectations, the front office can choose not to send documents to validation framework and process the complete batch automatically.

2.4.4 Configuration Manager

ICTP has a configuration service which reads the configuration properties and displays it in user interface to back office or IT people. The user can then change the configuration and the change is reflected immediately without restart of application server.

2.4.5 Available as API

Once required training is done, ICTP can be accessed through APIs. For example, if only classification of the document that needs to be known, then consumer application can call classification service and ICTP will return, classification results in JSON format.

2.4.6 Plug and Play framework

Using an orchestration module and configuration file, ICTP is implemented on plug and play framework. For example, ICTP has tesseract as OCR engine as its inbuilt OCR engine. An organization can already have invested in some other OCR engine. In this scenario, ICTP can use the OCR engine, purchased by the organization, and deliver desired result.

2.4.7 Light Weight

Although ICTP is rich in features, still it is light weight application and can be deployed as a WAR file.

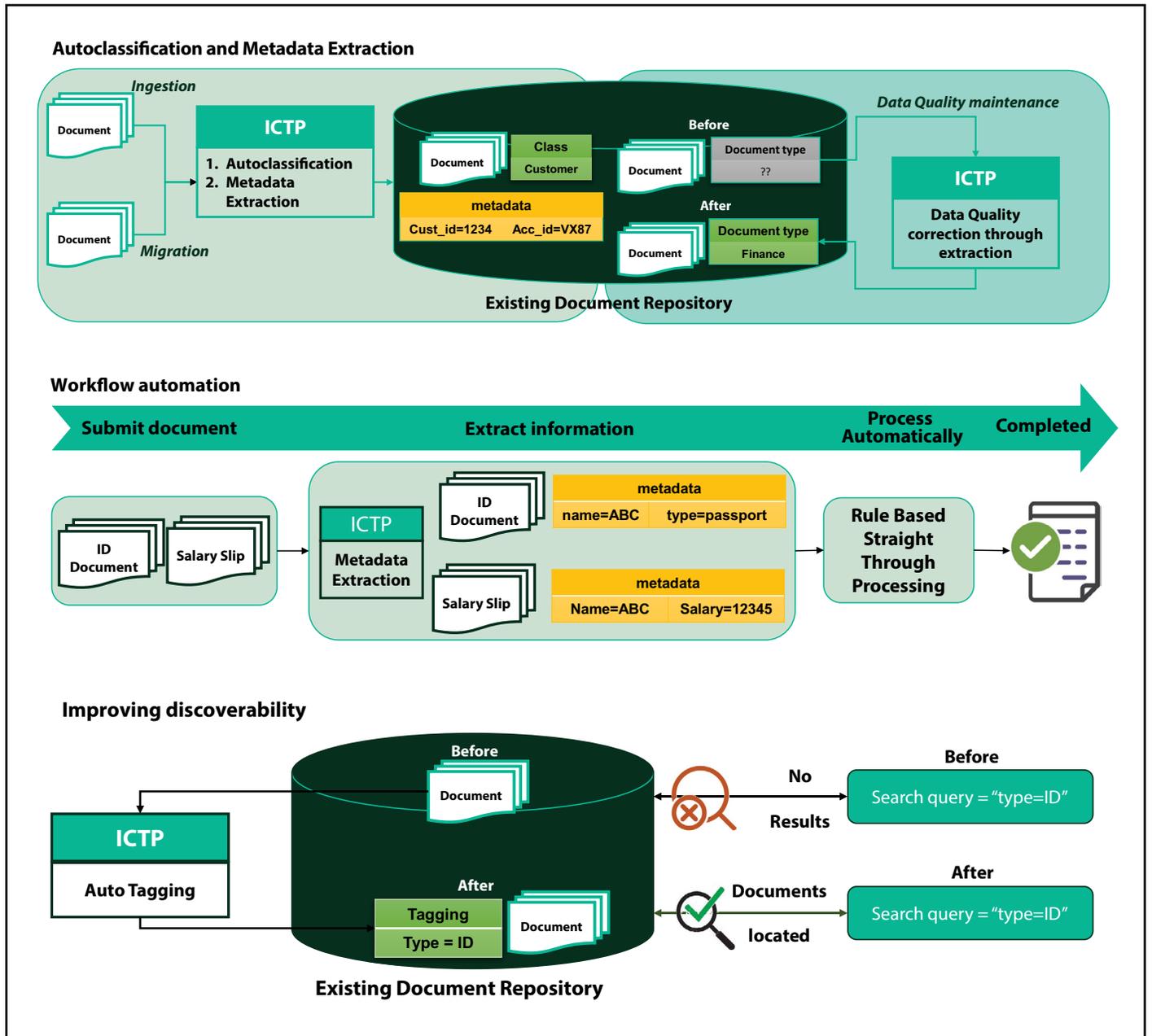
2.5 How ICTP can be used to integrate content intelligence

In this section some of the use cases, where ICTP can be used, are discussed. It should be noted that, in all following use cases, multimedia files such as audio and video can also be used. The audio and video files need to be converted to text using some “Speech to Text” (for audio file) tool. ICTP can use this text for further processing.

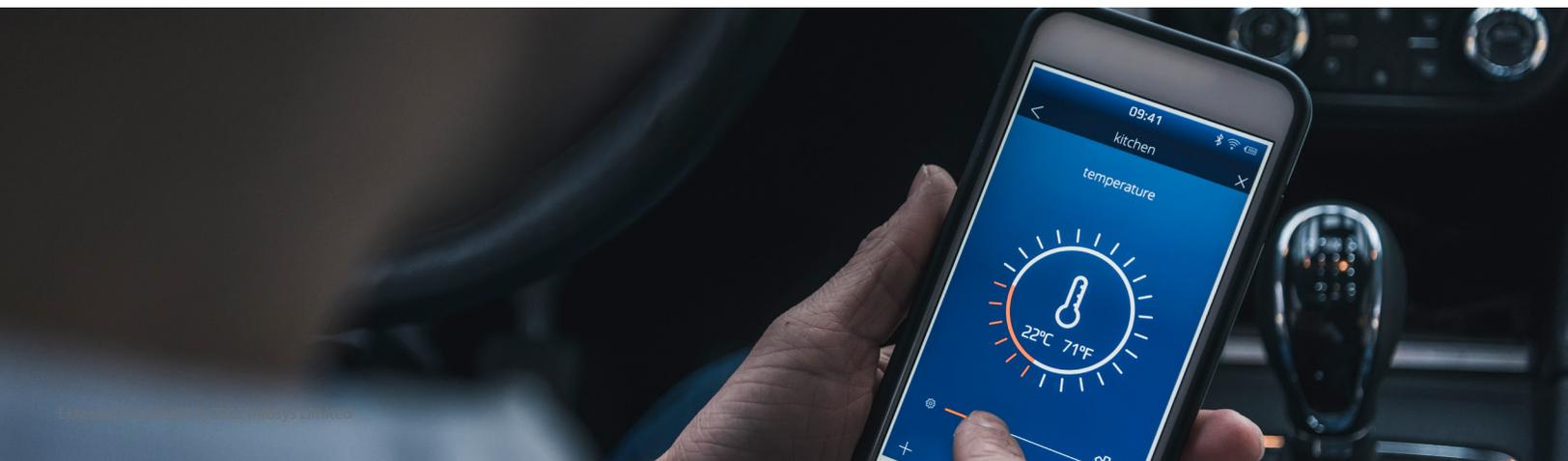
Use case	Description	Sample Scenarios
Auto Classification	ICTP can classify a document into a category based on occurrence of a keyword	<ul style="list-style-type: none"> • Document Type Identification. ICTP can identify the type of document through classification. • Classification of documents based on taxonomy. Alternatively, ICTP can also be used to check whether the current classification of document is correct or not. • Classification of a contract or agreement as risky, based on occurrence of aggressive terms used in the contract or agreement
Metadata Extraction	ICTP can extract contextual metadata using Named Entity Recognition (NER)	<ul style="list-style-type: none"> • Extracting metadata from ID documents. ICTP can extract the metadata from ID documents and then the application can display these values on user interface. • Migration. ICTP can be used in migration of documents from sources, where there is no metadata, such as shared drives, etc. to content services repositories. • Identity or locate documents with PI information for GDPR compliance. ICTP has inbuilt AI models which can identify known entities such as person, places, and dates. This makes ICTP capable of identifying the PI information contained in a document.
Workflow Automation	ICTP can automatically process the information to automate document workflow	<ul style="list-style-type: none"> • Automating loan processing workflow. ICTP can extract customer details, loan information etc. and provide the workflow engine which can then automatically process the application • Automating KYC. ICTP can validate KYC documents, extract appropriate information and can-do basic validation such as customer name is same on all documents, passport expiry date is not a past date, etc. and complete KYC process.
Data Quality Maintenance	ICTP can maintain quality of metadata	<ul style="list-style-type: none"> • Verify, validate, or enrich metadata. ICTP can pull documents, analyses the document, and extract the contextual metadata from the documents and tag the metadata to the document or verify metadata’s correctness.
Content Monetization by improving discoverability	ICTP can improve document discoverability	<ul style="list-style-type: none"> • Tagging document by additional metadata. ICTP can identify important business specific terms in a document and can tag it to the document as additional metadata. With this additional tagging, the document can be located quickly.
Summarization	ICTP can create summary of a large document	<ul style="list-style-type: none"> • Document Summery. ICTP can create summary of a large document by identifying important business terms.
End to End Digitalization	ICTP can be used for end-to-end digitalization of physical records	<ul style="list-style-type: none"> • 100% Digitalization. ICTP can extract contextual metadata information from scanned documents and can tag it to the document while saving documents into content services repository. Documents that cannot be processed, are sent to full text search engine, where such documents can be located using full text engine.
Content Generation	ICTP can generate the content for agreement, communication etc.	<ul style="list-style-type: none"> • ICTP can generate personalized content for communication as well as for contracts & agreements. Based

2.6 How ICTP fit into existing landscape

The following diagram shows how ICTP can fit into existing landscape



This ICTP creates a Content Intelligence layer which can be leveraged for tagging, workflow automation, improving discoverability etc. which is represented as follows.

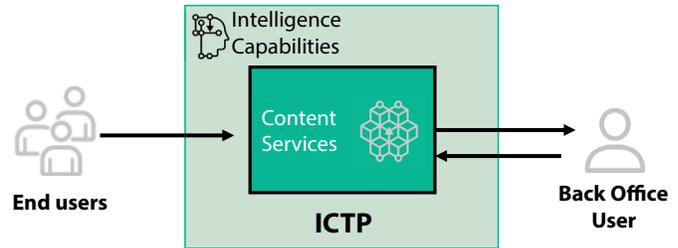


Automation powered by ICTP



Without ICTP

- Intelligence - Manual by Backoffice User
- Classification - Manual
- Metadata Extraction – Manual
- Workflow Automation – Partially automated
- Data Quality Maintenance – Manual
- Content Monetization – Manual
- Summarization - Manual
- Digitalization of Data – Partially automated



With ICTP

- Intelligence - Automated
- Classification - Automated
- Metadata Extraction – Automated
- Workflow – Automated
- Data Quality Maintenance – Automated
- Content Monetization – Automated
- Summarization - Automated
- Digitalization of Data – Automated

To automate manual tasks cognitive capabilities are required. ICTP creates a Content Intelligent layer which then can be consumed to take business related decision and automate the task.

3 Pros and Cons of ICTP

The following table list Pros and Cons ICTP

Pros	Cons
<ul style="list-style-type: none"> • ICTP can be leveraged for offering content intelligent service along with existing Content Services tools as add-on or as complementary AI capability. • ICTP can be trained using excel based templates which reduces training time from months to weeks. • Because of its easy to use and intuitive user interface, business user can easily work with ICTP. • ICTP artificially generates the data, which reduces the amount of “training data” needed for training. • Recommendation engine helps business user to identify additional important terms. • Business user can verify the processing result, make changes, and then complete the processing. • Business users can also provide feedback to back-office users using the inbuilt feedback loop. • ICTP is built on plug and play architecture. A new component such as OCR engine can be added into the processing pipeline. • ICTP supports structured, semi-structured and unstructured documents 	<ul style="list-style-type: none"> • For more advance OCR capabilities such as ICR, advance OCR engine is needed. Based on requirement new OCR engine can be integrated. • Classification and extraction is a step-by-step process. For Initial set of documents, a manual user verification is recommended to measure the accuracy and then push the documents to the next steps. Once accuracy reached to an acceptable level, this verification can be bypassed. • Basic image enhancement is available. For advance image enhancement capabilities, additional image enhancement module needs to be added. • In ICTP, position-based extraction is not preferred. Hence a semantic, syntax, relationship-based approach is used for entity extraction.

4 Conclusion

ICTP is built using a bottom-up approach where core capability is classification and extraction. With this core capability, ICTP is a versatile solution which can be leveraged for many use cases where content intelligence is needed. In Content Services, ICTP can be used to fulfill the following sample use cases

Content Lifecycle Stages				
CAPTURE & INGESTION	MANAGE	DISTRIBUTE	SEARCH	PRESERVE
				
ICTP Usage scenarios				
Auto classification based on occurrence of terms and phrases	Create summary of content which can be used understand complete content	Generate personalized content for communication, agreements & contracts	Extract and tag the content with entities in the content	PI information can be identified and set to retention, different ACL etc.
Automatically extract contextual metadata during migration or ingestion	Verify and validate content metadata	Validate the politeness of language in outgoing correspondence	Query enrichment to add additional relevant criteria for better search result	Based on occurrence of terms, content can be classified as confidential
Migrate legacy systems such as Shared Drive to Content Services by extracting metadata	Provide appropriate business attributes to automate processing of workflow	Identify the risk in agreement or contract	Understanding the context and providing search results based on the context	Erich content metadata based on new requirements

Some of the key implementations are automatic metadata tagging and classification, Content intelligence driven workflow automation, improving the document discoverability, summarization, and end to end digitalization. With proper plug-in, ICTP can process multimedia audio and video files and can deliver the same result.

ICTP is designed keeping in view, "what is needed at ground level", it stands out significantly pretty well as compared to other products in the same space.

About the Authors



Girish Pande
Principle Technology Architect



Yamuna Kannaian
Senior Technology Architect



Anoop Kumar P
AVP, Senior Principal
Technology Architect



About the Mentors

For more information, contact askus@infosys.com



© 2022 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.