Every organization processes documents in various formats, especially paper and fax formats. Processing documents manually is an expensive and time-consuming endeavor. Ephesoft Enterprise is a modern document capture solution that allows an organization to automate the business process. It uses powerful technology to classify and capture the vital information from the document's content. This helps to minimize the time spent on reviewing and processing any documents.

Start by learning about document capture and how Ephesoft revolutionized the industry. Progress to a tour of key features, including the operator and administrator interfaces and then learn to configure Ephesoft to process your business’s specific document types and extract content from those documents. You will also get to know the advanced customization techniques that make Ephesoft accommodate your unique business needs. Finally, the book concludes by teaching you how to embed the classification and extraction functionalities using Ephesoft’s web services.

Who this book is written for
This book is intended for IT professionals interested in installing and configuring Ephesoft Enterprise for their organization, but it is a valuable resource for anyone interested in learning about intelligent document capture.
In this package, you will find:

- The author's biography
- A preview chapter from the book, Chapter 1 'A Quick Tour of Ephesoft'
- A synopsis of the book’s content
About the Authors

**Pat Myers** is the executive vice president and a cofounder of Zia Consulting, a content-centric solutions firm. Zia is a platinum Ephesoft and Alfresco partner that provides solutions from paper to mobile. Zia was the Ephesoft Partner of the Year in 2012, 2013, and 2014. Pat has over 14 years of Enterprise Content Management (ECM) experience and 18 years of experience in professional services and application development. Pat and Ike Kavas developed the first official Ephesoft training, and Pat is the coauthor of *Intelligent Document Capture with Ephesoft, First Edition*.

**Ike Kavas** has more than 15 years of solid experience in document imaging, document management, workflow, and systems. He is the founder and the chief technology officer at Ephesoft, Inc., responsible for product design and road maps. He has a keen technical background, which he has developed by implementing several multimillion-dollar projects for Fortune 100 companies and has outstanding sales and business experience, which he has demonstrated by achieving and exceeding revenue-based goals.

Kavas is a serial entrepreneur. Before founding Ephesoft, Inc., he managed professional services at Kofax, Inc. and cofounded other technology companies in southern California. He holds a bachelor's of science degree in electronics and electrical engineering as well as a CDIA+ certification.

**Michael Muller** is the director of engineering at Zia Consulting. He has 25 years of professional software development experience and is currently specializing in enterprise content management.

**Jon Solove** is a senior solutions engineer and ECM consultant at Zia Consulting, a content-centric solutions firm. He designs, implements, and sells Capture and ECM solutions for a multitude of industries, from manufacturing to financial services. He is a platinum Ephesoft and Alfresco partner who provides solutions from paper to mobile. Zia has been Alfresco's Partner of the Year in 2012, 2013, and 2015 and Ephesoft's Partner of the Year in 2012, 2013, and 2014.
Preface

In general, document capture refers to the process of scanning paper documents using scanners or cameras, and transforming these documents into an electronic file such as a PDF, TIFF, and so on. Through a type of document capture software, these electronic files are assessed through character or pattern recognition (that is, OCR, ICR, and OMR) and converted into meaningful data or information, also called metadata. The goal of document capture systems is to classify incoming documents into categories and extract metadata by automating processes that humans normally do. By automating this process, organizations can classify documents and route the incoming documents to repositories and workflow systems with all the metadata more efficiently; thus, document capture systems help reduce errors, allow documents to be handled faster, and organizations can scale their business using software rather than labor.

Over the years, the document capture industry has evolved in many ways. Paper documents are still a big portion of the document formats that organizations receive. However, as organizations started to exchange documents electronically, the document capture systems had to evolve to be able to process the documents within e-mail attachments, sent via FTP or by using APIs. Of course, smartphone adoption also influenced the document capture world. As consumers started to produce electronic documents more using their phones, such as check deposits or expense reports, capture using mobile devices became essential.
The capabilities of the capture software has also become more intelligent over the years. Early systems automated the capture process by using barcodes for classifying documents and extracting metadata from predefined areas of structured documents, called forms. Later on, technologies such as Ephesoft classified the documents using several techniques such as document layout or words and phrases that can extract the metadata without defining where the metadata might be located on the document. We call these capture systems Intelligent Document Capture systems and they can help organizations automate document capture systems not only for structured documents but also for unstructured documents such as e-mails or documents that do not have any known format. Intelligent Document Capture systems are also easier to deploy and maintain compared to older systems, which provides faster ROI and adoption.

The latest trend in document capture systems is the use of web-based APIs. In the beginning, document capture systems were used only to ingest documents to repositories and workflow systems but with the popularity of APIs and the cloud, a new use case has opened. The newest versions of document capture systems now allow organizations to enable other business applications with document capture functionalities, where the cloud, private or public, based APIs provide document capture services rather than the applications. This makes organizations more efficient when it comes to maintaining, upgrading, and integrating multiple business applications.

If we try to imagine where document capture will be 5 to 10 years from now, it is anyone's guess. However, looking at the emerging technologies and today's advancements, we may be able to predict the future. In the future, we will see machine learning algorithms doing what capture administrators do today, so that any user can simply show the next generation document capture system what he/she wants within that document. Then, the computer should be able to understand what to do the next time without any human intervention. What this means is, if it takes one administrator to configure the system today and one IT professional to set up the servers, in the future there will be no need for either resource. The users will simply request a capture service using a browser or mobile device, the service will learn the human behavior by simply observing the user and then will perform the same task from then on. The projects on which we used to spend months on implementing are now measured in weeks. In the future, they will be measured in hours, if not minutes and seconds.
What this book covers

Chapter 1, A Quick Tour of Ephesoft, takes you on a walkthrough of Ephesoft's user interface. We will look at the administrative and the operator functionality of Ephesoft.

Chapter 2, Creating a Batch Class, explains how to set up Ephesoft. We will see how to create a new batch class and configure it for classification and extraction.

Chapter 3, Core Ephesoft Features, expands on the features introduced in Chapter 2, Creating a Batch Class. We will learn more about classification and indexing techniques. We will also learn about web scanning.

Chapter 4, Ephesoft's Advanced Features, explains advanced Ephesoft features. This includes how to set up Microsoft Active Directory, scripting, and utilizing web services.

Chapter 5, Tips, covers the productivity-enhancing tips.

Appendix, References, includes some reference material.
As an introduction to Ephesoft, we will first walk you through the user interface and then examine the installation folder. The locations of certain files and folders within the Ephesoft installation are important because an administrator must make changes here to enable some features.

In this chapter, we will examine the following aspects of Ephesoft:

- The user interface
- The installation folder

The user interface

After logging in, users can access Ephesoft's features from an automatically hiding menu of navigation items that we will refer to as the side navigation. To display this menu, simply move your mouse cursor to the left-hand side of the browser window.

Ephesoft has organized this side navigation so that administrative features are separate from the common functions that operators use. Operators typically submit batches and review and validate Ephesoft's output, supplying additional information about the document images being processed.
Administrators enable these activities by defining the operations to be performed on each type of batch. Administrators also monitor and control the processing of the batches.

**Administrative features**

The side navigation provides links to five areas of the system that are commonly used by administrators:

- Batch class management
- Batch instance management
- Folder management
- System configuration
- Reports

**Batch class management**

A batch class defines a set of operations that should be performed on the page images that are provided as input. A batch class consists of document types, document fields, batch class fields, e-mail configuration, and workflow/plugin configuration. The Batch Class Management interface allows administrators to create, modify, edit, and delete batch classes.
The batch class management interface displays a list of batch classes. Administrators can open a batch class to configure the following:

- **Document types**: The documents that will be processed in the batch class are configured here. Each document type is described by a distinct set of properties called fields. Rules can be configured to extract information from the document into the fields, thereby automating the process of indexing the document.

- **Modules**: Modules are the major steps in the processing of documents. Each module is implemented by a series of plugins.

- **E-mail configuration**: In this portion of the administrative interface, users can provide connection information for an e-mail account, and Ephesoft will process e-mails sent to this address. Ephesoft processes both the e-mail body and the attached documents.

- **Scanner profiles**: This is where administrators can associate one or more scanner configurations with each batch class. These profiles are available in the web scanner.

- **CMIS import**: CMIS is a standard protocol for communicating with document repositories. Ephesoft can use CMIS to monitor a standards-compliant document repository for input.
**Batch class fields**: Ephesoft can associate information with a batch (the group of page images that are processed together) as a whole. Each piece of information associated with a batch is called a batch class field. Batch class fields are applied to batches and should not be confused with document fields, which contain information that applies to individual documents.

---

**Batch instance management**

A *batch instance* is a set of page images processed together. The terms batch and batch instance are usually interchangeable. This area within the administrative interface allows administrators to see the status of batches, reprioritize batches, and restart batches in a previous processing step.

---

Ephesoft's batch instance management user interface
Folder management

The folder management interface allows the administrator to upload files for batch class configuration. These files are also accessible from the installation folder, but this is often a more convenient way to manipulate these files.

Ephesoft's folder management user interface
System configuration

This administrative interface allows users to manage Ephesoft in ways that are not specific to a batch class or instance.

System configuration allows the modification of the following features:

- **Regex pool**: The regular expression pool is a library of regular expressions that administrators can access when creating extraction rules for a batch class.

- **Workflow management**: Ephesoft's features are implemented in components called plugins. The workflow is the sequence in which these plugins are executed. This portion of the user interface allows an administrator to specify what plugins are available when configuring the workflow for a batch class.

- **Connection manager**: The connection manager allows you to create and test database connections. These connections are used by plugins to access databases.

- **License details**: This allows administrators to see the expiration date of the license and the features that are enabled.
Reports

Reporting can be enabled to provide administrators with statistics on the system and throughput. The administrator can filter reports by criteria such as batch class or start date. Advanced reports are also available, including correction reporting. Correction reporting identifies when operators made corrections to Ephesoft's automated processing. This information can be used to optimize the configuration over time.

The operator user interface

The side navigation provides links to the following four areas of the system that are commonly used by operators:

- Batch list
- Review validate
- Web scanner
- Upload batch
**Batch list**

The **batch list** shows the batch instances that require review or validation.

The review process involves documents that could not be identified as being of a certain type. In Ephesoft, as with most image capture systems, we say that these documents could not be classified. The review interface allows operators to split and merge pages of documents and specify the document type.

The validation process involves fields for which data could not be extracted from the document, or fields where the extracted values do not comply with the previously specified rules.

![Ephesoft's batch list user interface](image)
Review validate

The review validate screen will present the operator with the next available batch for processing according to priority and batch date.
Web scanner

Ephesoft is capable of capturing content from a scanner attached to the user's workstation. What is unique about the web scanner is that no software needs to be installed on the workstation; Ephesoft uses a Java applet to send content directly to the server from any TWAIN-enabled scanner.

![Ephesoft's web scanner user interface](image)

The first time a user logs into the operator interface and selects the **Web Scanner** link on the side navigation, the user will have to choose a scanner. When the user selects the **Source** button, the user will be shown all TWAIN devices that have been installed on the user's workstation. Once the scanner is selected, the user can select the batch class to be used for processing and start the scan job.

Upload batch

Operators can submit PDF and TIF files directly to Ephesoft for processing by using the upload batch feature. Once the documents are selected and uploaded, the operator can select the appropriate batch class and start the batch processing.
File system

The following are some important folders that are created when Ephesoft is installed. These are subfolders beneath the Ephesoft installation folder:

- **Apache 2.2:** Apache can be used in front of Ephesoft for load balancing and failover. It is included in the installation but not configured.
- **Application:** The Ephesoft Java web application is installed in this folder.
- **Application/i18n, themes:** These folders contain files to customize and localize the Ephesoft application.
- **Application/native/RecostarPlugin:** This plugin provides the image OCR functionality.
- **Application/WEB-INF/classes/META-INF:** System configuration property files are stored in this folder.
- **Dependencies.gs, ImageMagick:** Applications that Ephesoft uses for image manipulation are installed here.
- **Dependencies/licence-util, licensing:** These folders contain tools to collect the information needed to generate and install license keys.
- **Dependencies/luke:** Luke is a tool that helps troubleshoot problems with Lucene indexes.
A Quick Tour of Ephesoft

- **JavaAppServer**: This folder contains the Tomcat configuration for Ephesoft.
- **JavaAppServer/conf**: This is where the contexts are defined for Ephesoft; it is where URLs are bound to java code.
- **EphesoftReports**: The configuration and binaries for reporting are stored here.
- **SharedFolders/BC99**: The configuration for each batch class is stored here. The contents of the batch class folder can be modified through the **Folder Management** interface by a batch class or system administrator.

**Summary**

In this chapter, we looked at the administrative and the operator functionality of Ephesoft. We also looked at the installation folder on the filesystem. It's time to put Ephesoft to work.

In the next chapter, you'll learn how to train the system to recognize your documents, extract content from them, and test the configuration.
Where to buy this book


Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.

Click here for ordering and shipping details.