



IRISXtract™ for Documents (X4D)

I.R.I.S. AG

Inhaltsverzeichnis

1. IRISXTRACT™ FOR DOCUMENTS (X4D)	3
1.1 WOM-SERVER-APPLIKATION.....	4
1.2 WOM-CONTROLLER	5
1.3 X4D-MANAGER	5
1.4 SOLUTION DESIGNER	5
1.5 IMPORT UND EXPORT-KOMPONENTEN	7
1.6 ANALYZE	7
1.7 LEISTUNGSFÄHIGE KLASSEKATIONSTECHNOLOGIE	7
1.8 VERIFY.....	9
1.9 WEBBASIERTES VERIFY	9
1.10 XTRACT FUZZY DB	10

Abbildungsverzeichnis

ABB.1 SYSTEMÜBERSICHT XTRACT FOR DOCUMENTS	4
ABB.2 MANAGEMENT-KOMPONENTEN.....	5
ABB.3 INTEGRIERTE VSTA-ENTWICKLUNGSUMGEBUNG.....	6
ABB.4 DER SOLUTION DESIGNER IN RELATION ZUR PLATTFORM	6
ABB.5 XCLASSIFY-STRUKTUR	7
ABB.6 VERIFY.....	9
ABB.7 VERIFY.RDP.....	9
ABB.8 VERIFY-QUEUES.....	9
ABB.9 WEBBASIERTE REMOTE VERIFY-ANWENDUNG	10

1. IRISXtract™ for Documents (X4D)

IRISXtract™ for Documents ist eine skalierbare Client/Server-Applikation für Microsoft-Betriebssysteme, die mit Hilfe optischer Zeichenerkennung und Bildverarbeitungsalgorithmen Dokumente grafisch und textlich klassifizieren und ausgewählte textliche Informationen aus digitalisierten Dokumenten extrahieren kann. Digitale Dokumente werden üblicherweise mit Hilfe eines Multifunktionsgerätes oder Dokumentenscanners erzeugt, alternativ könnten aber auch andere elektronische Daten, die zuvor mit Hilfe externer Datenkonverter in ein bildhaftes Format umgewandelt wurden, verarbeitet werden.

I.R.I.S.-Lösungen auf Basis der „all-in-one“- Plattform IRISXtract™ for Documents optimieren sämtliche unternehmenskritischen Business-Prozesse. Das System bietet die Möglichkeit, beliebige Dokumentenarten in einem System vollständig zu verarbeiten - angefangen beim Datenimport über die Klassifizierung von Eingangspost und anschließender Datenextraktion bis hin zur Weiterleitung der Daten für die abschließende Verarbeitung der erfassten Daten im zentralen EDV-, ERP- oder CRM-System.

IRISXtract™ for Documents wird für unterschiedliche Zwecke erfolgreich eingesetzt. Anwendungen sind beispielsweise die Datenerfassung für eingehende kreditorische Rechnungen, die Datenerfassung für Formulare im Gesundheits- und Bankenwesen oder die Klassifikation von Eingangspost in Digital Mailroom Lösungen oder die Verarbeitung von elektronischen Akten für unterschiedliche vertikale Märkte. IRISXtract™ for Documents wird seit Jahren im Dienstleistungsumfeld sehr erfolgreich eingesetzt, weil die zugrundeliegende Architektur auch sehr professionellen Ansprüchen genügt.

IRISXtract™ for Documents kann den gesamten Ablauf während des Extraktionsprozesses (Datenimport, automatische Datenextraktion, manuelle Prüfung bzw. Korrektur der Extraktionsergebnisse, Datenexport) kontrollieren. In diesem Sinne besitzt IRISXtract for Documents Workflow-Management-Funktionalität.

IRISXtract™ for Documents arbeitet mit Dokumentenstapeln und ist typischerweise in eine Prozesskette eingebunden, d.h. das System empfängt zu verarbeitende Bilddaten von einer Datenquelle (Bilddateien von Scanner, Fax-Server, E-Mail-Server mit Bildkonvertierung, Datenbanken, etc.) und liefert Ausgabedaten typischerweise in maschinenlesbarer Form (XML- und Textdateien, PDF, PDF/A, etc.) an nachfolgende Systeme zur fachlichen Weiterverarbeitung (ERP-Systeme, Archive, Datenbanken, etc.).

In der folgenden Beschreibung möchten wir den grundsätzlichen Aufbau sowie die prinzipiellen Eigenschaften des IRISXtract™ for Documents-Systems erläutern. Weitergehende Erklärungen einzelner Funktionen und detaillierte Beschreibungen der Systemkomponenten können der entsprechenden Systemdokumentation entnommen werden.

IRISXtract™ for Documents ist ein modulares System und als skalierbare Client-Server-Architektur realisiert, die höchste Transaktionssicherheit im Produktivbetrieb gewährleistet. Die Kommunikation zwischen WOM-Server und Clients (z.B. IMPORT, ANALYZE, VERIFY, EXPORT) erfolgt über Microsoft COM+, Microsoft Message Queue, OLE DB, TCP/IP, UDP und über direkten Dateizugriff. Die Skalierung bietet eine enorme Bandbreite, ohne dabei die grundlegende Architektur verändern zu müssen: Applikationen mit kleinem Durchsatz können auf lediglich einer Maschine implementiert werden, hochvolumige Anwendungen können auf mehrere hundert Maschinen verteilt sein.

System-Konfigurationen sind skalierungsunabhängig, damit können ohne Probleme zwischen Entwicklungs-, Test- und Produktsystemen Konfigurationsdaten ausgetauscht werden.

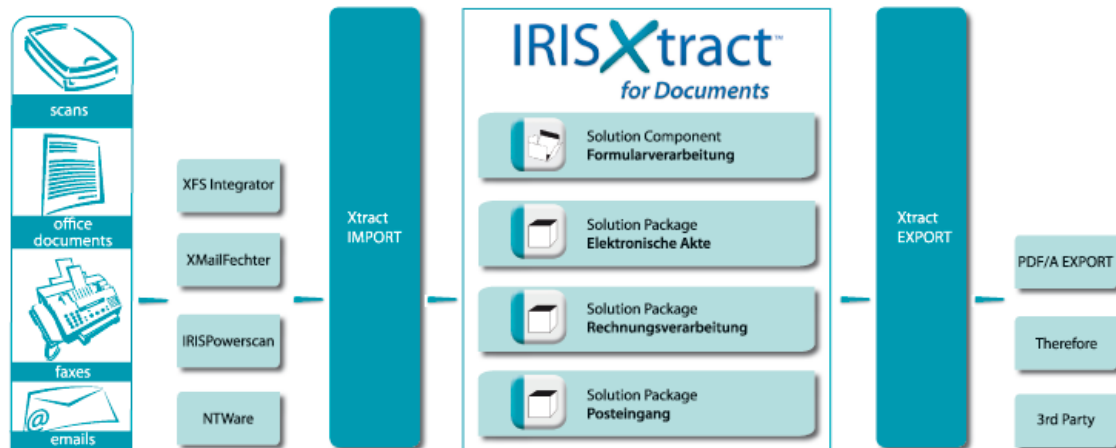


Abb./Tab. 1: Systemübersicht Xtract for Documents

Die Skalierung und Lastverteilung innerhalb des Systems erfolgt auf Basis eines Warteschlangenkonzeptes: serverseitig werden Jobs in Warteschlangen eingestellt, je nach Durchsatzanforderung werden die Jobs von 1 bis n parallel laufenden Clients angenommen, für die Dauer der Verarbeitung gesperrt, bearbeitet und anschließend wieder zurückgegeben. Die maximale Anzahl der gleichzeitig mit einem WOM-Server verbundenen X4D-Clients ist abhängig von der eingesetzten Hardware und nicht durch die Software begrenzt.

IRISXtract™ for Documents kann als mandantenfähiges System betrieben werden. Dabei werden mehrere Projektkonfigurationen zeitgleich und unabhängig voneinander im System gehalten und auf unterschiedliche Dokumentenströme angewendet. Aus Operating-Sicht verhält sich IRISXtract™ for Documents wie ein virtualisierbares System.

Die X4D-Plattform besteht aus folgenden Komponenten:

1.1 WOM-Server-Applikation

Die WOM-Server-Applikation bildet die Schnittstelle zwischen der Prozessdatenbank (Microsoft SQL Server), dem Microsoft Message Queue Service und den Xtract-Clients. Alle Clients kommunizieren mit der Server-Applikation über COM+. Die WOM-Server-Applikation kapselt den Betrieb der Prozessdatenbank (Kommunikation über OLE DB) und den Betrieb des Microsoft Message Queue Services (Kommunikation über COM+).

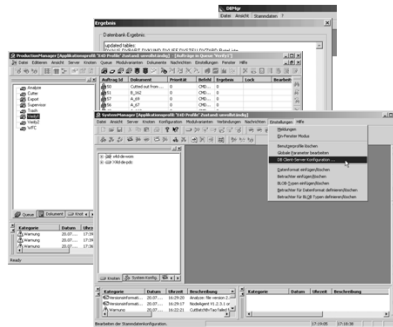


Abb./Tab. 2: Management-Komponenten

1.2 WOM-Controller

Der WOM-Controller steuert den Erfassungsworkflow, indem er den von den Verarbeitungs-Clients (Import, ANALYZE, VERIFY, Export) zurückgelieferten Dokumentenstapeln anhand der Dokumentenstati neue Verarbeitungsaufträge zuweist (Beispiel: Nach dem Import wird ein Dokumentenstapel in die Analyse-Warteschlange eingestellt).

1.3 X4D-Manager

Der X4D-Manager für die Produktion und das System erlaubt die Überwachung und Steuerung der eigentlichen Dokumentenverarbeitung im X4D-System und dessen Komponenten.

1.4 Solution Designer

Der Solution Designer besitzt Workbenches für die Konfiguration und das Customizing von Formular- und Freiformlösungen für Klassifikation und Indizierung. Die Workbenches bieten umfangreiche grafische Benutzeroberflächen zum Editieren der Projektkonfiguration, die mit Hilfe des Produktions-Managers zentral in das IRISXtract™ for Documents-System eingespielt und anschließend automatisch verteilt wird. IRISXtract™ for Documents veröffentlicht ein sehr umfangreiches Objekt- und Event-Modell, das nahezu vollständig den Zugriff auf alle relevanten Eigenschaften der Produktionsdaten und des Systems selbst ermöglicht. Zwei in den Solution Designer integrierte Entwicklungsumgebungen auf Basis von Visual Basic Script (VBS) und insbesondere Visual Studio Tools for Applications (VSTA) greifen auf dieses Objekt- und Eventmodell zu und ermöglichen ein sehr leistungsfähiges Customizing der gesamten Produktivstrecke.

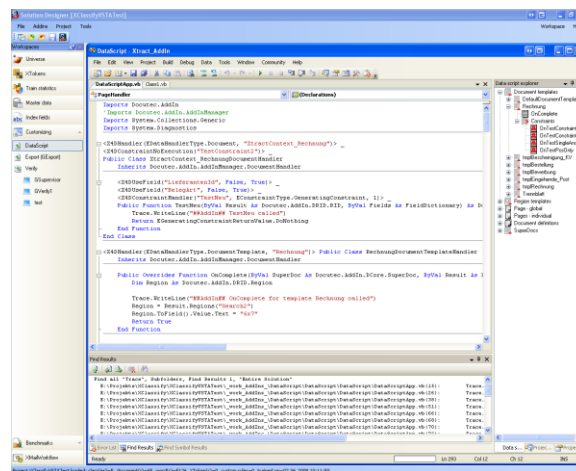


Abb./Tab. 3: Integrierte VSTA-Entwicklungsumgebung

VSTA ist unmittelbar aus dem Microsoft Developer Studio abgeleitet und kann durchaus als das leistungsfähigste im Markt erhältliche Customizing-Framework betrachtet werden. Basierend auf .Net-Technologie bietet VSTA hervorragende technische Möglichkeiten für Entwickler, die bisher nur professionellen Entwicklungsumgebungen vorbehalten waren: die umfangreichen .Net-Klassenbibliotheken vereinfachen und erweitern die objektorientierte X4D-Anwendungsentwicklung dramatisch. Funktionen wie Intellisense, Codefolding, die integrierte Hilfe, Live-Debugging oder das zentralisierte Deployment bilden die Grundlage für kurze Entwicklungs- und Testzyklen. Die .Net-Technologie ermöglicht die einfache Integration von Fremdkomponenten, dabei werden alle relevanten Schnittstellenstandards (COM, ActiveX, .Net) unterstützt. VSTA bildet damit auch eine hervorragende Basis für die effiziente und leistungsfähige Kopplung an externe Systeme und Datenbanken.

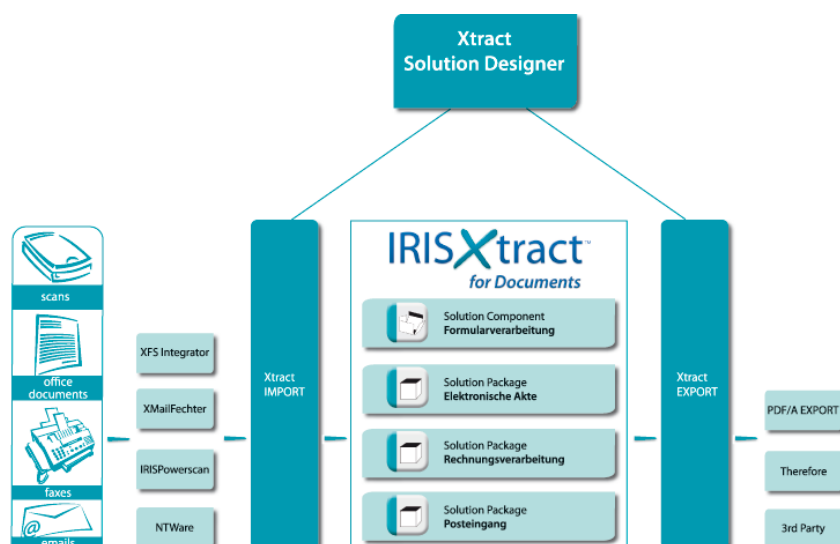


Abb./Tab. 4: Der Solution Designer in Relation zur Plattform

1.5 IMPORT und EXPORT-Komponenten

Die IMPORT-Komponenten speichern zu verarbeitende Dokumente als Bilder und optionale Metainformationen in die WOM-Datenbank. Die EXPORT-Komponenten leiten Bilder und Indexdaten an nachfolgende Systeme weiter. Beide Komponenten besitzen Schnittstellen für standardisierten Datenaustausch, darüber hinaus kann das Verhalten der Module über VBS- oder VSTA-Komponentenskripte nahezu unbegrenzt angepasst werden. Die Kopplung an beliebige Fremdsysteme mittels überwachter Prozesse ist in der Regel problemlos realisierbar.

1.6 ANALYZE

Die ANALYZE-Komponenten klassifizieren und indizieren Dokumente, extrahieren die gesuchten Informationen aus den Dokumenten und prüfen abschließend die Ergebnisse. Der Analyseprozess läuft typischerweise in mehreren Schritten ab: (a) Verbesserung der Bildqualität. (b) Wenn möglich bildhafte Klassifikation anhand zuvor definierter Templates (XFingerPrint-Verfahren). (c) Alternativ und/oder ergänzend zum XFingerPrint-Verfahren Anwendung eines generischen Extraktionsregelwerkes auf den Freitext (XContext-Verfahren). (d) Generierung des logischen Dokumentes. (e) Prüfung des logischen Dokumentes (Datenskripte, Stammdatenabgleich, etc.).

1.7 Leistungsfähige Klassifikationstechnologie

IRISxtract™ for Documents Version 3.x etabliert einen neuen funktionalen Kern der I.R.I.S. Digital Mailroom Lösungen, nämlich „XClassify“. Die in XClassify verwendeten Methoden für die Dokumentenklassifikation basieren auf Dokumentenklassen, die lediglich über die Bereitstellung entsprechender Beispieldokumente definiert werden. Der technisch wesentliche Prozess innerhalb von XClassify besteht darin, in einem Trainingsdurchlauf automatisch generelle und effiziente interne Repräsentationen zu berechnen, die den statistischen Eigenschaften der natürlichsprachigen Beispieldokumente entsprechen. Dieser Prozess ist entscheidend für die Qualität der Klassifikationsergebnisse. Bei diesem Ansatz werden manuelle Administrationsaufwände auf ein Minimum reduziert: in der Regel muss lediglich die Zuordnung von Beispieldokumenten zu den korrekten Klassen geprüft und einige wenige globale Parameter eingestellt werden.

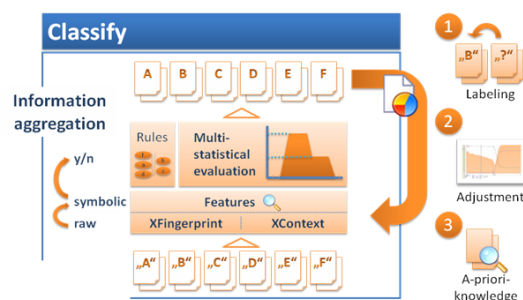


Abb./Tab. 5: XClassify-Struktur

1.8 VERIFY

Die VERIFY-Komponente ermöglicht entsprechend der Konfiguration die manuelle Überprüfung und Korrektur der vom Extraktionsprozess gelieferten Informationen. VERIFY zeigt die extrahierten Daten zusammen mit den zugehörigen Lesezonen und Bildern des verarbeiteten Dokumentes an. Die Daten können überprüft und korrigiert werden. IRISXtract™ for Documents unterstützt die Anzeige von Grauwert- oder Farbbildern für die manuelle Datenprüfung im VERIFY-Modul. Es können beliebige VERIFY-Rollen mit jeweils unterschiedlichem Verhalten definiert und im Workflow integriert werden (etwa Nacherfassung, Teamleiter, Trainer, Co-Trainer, Supervisor, etc.). Für die Anpassung an besondere Nacherfassungssituationen verfügt VERIFY über umfangreiche Customizing-Funktionen auf Basis von VBS und VSTA.

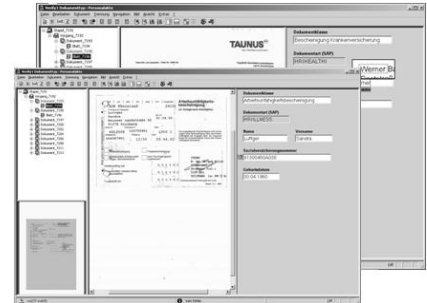


Abb./Tab. 6: Verify

1.9 Webbasiertes VERIFY

Bei der webbasierten Verifikation von Dokumenten steht 100%ige Kompatibilität der X4D-VERIFY Funktionalität über allen Anwendungs- und Implementierungsformen. Gerade in der weltweiten webbasierten Verwendung von VERIFY hat die I.R.I.S. AG besonderen Wert auf eine durchgängige Integration von IRISXtract™ for Documents gelegt. Über die Standard Virtualisierungsfunktionen (etwa HyperV) der Windows Betriebssysteme ab Server 2008 können die IRISXtract™ for Documents Module VERIFY und Production-Manager vollständig weltweit im WEB veröffentlicht werden.

Auf Basis der individuellen Berechtigungen wird von jedem Rechner auf der Welt eine direkte Verbindung zu dem veröffentlichten X4D-Server aufgebaut. Diese Verbindung ist modulspezifisch, hier ist die VERIFY-Anwendung zu sehen.

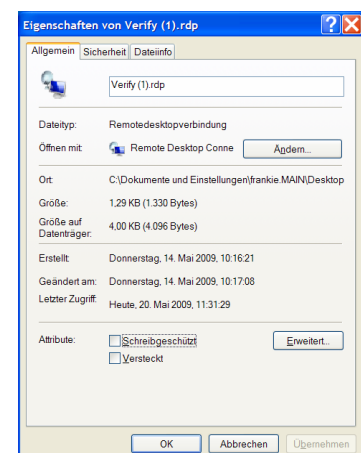


Abb./Tab. 7: Verify.rdp

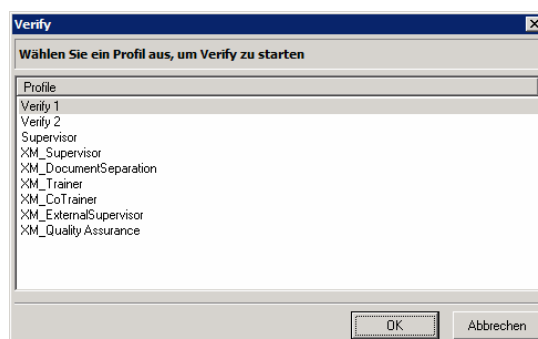
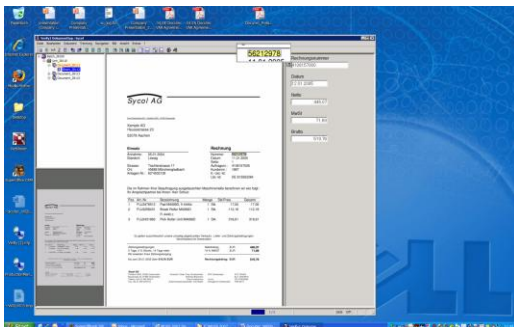


Abb./Tab. 8: VERIFY-Queues

Nach Durchführung der Berechtigungsverfahren wird dem Anwender der definierte Zugriff auf die VERIFY-Queues gewährt.

Anschließend kann der Anwender VERIFY remote über das WEB starten.

Wie man an der folgenden Darstellung sehen kann, läuft das VERIFY auf dem remote-Desktop, ohne dass irgendwo ein Xtract-WOM System läuft:



Dem Anwender stehen dadurch alle VERIFY-Funktionen zur Verfügung. Es existieren keinerlei Einschränkungen, die webbasierte VERIFY-Anwendung steht überall auf der Welt in voller Bandbreite zur Verfügung, nur drei Klicks entfernt! Ausschließlich die Bandbreite der Webverbindung an sich ist hier der limitierende Faktor.

Abb./Tab. 9: Webbasierte remote Verify-Anwendung

1.10 Xtract Fuzzy DB

IRISXtract™ for Documents arbeitet mit kompilierten Stammdaten, die extrem schnelle unscharfe Datenabgleiche auf Basis der Berechnung von Levenstein-Distanzen ermöglichen. Stammdaten werden in standardisierten ASCII-CSV-Formaten bereitgestellt, die Kompilierung der Stammdaten und die Einstellung/Verteilung in das X4D-System erfolgt durch den X4D-Manager. Darüber hinaus können Online-Datenabgleiche über Daten- und Komponentenskripting realisiert werden.

I.R.I.S. Kontaktinformation:

Heussstraße 23

52078 Aachen

Germany

Tel: +49-(0)241-92035-0

Fax: +49-(0)241-92035-25

E-Mail: info-de@iriscompany.com

© Copyright 2010 I.R.I.S. s.a./n.v./AG

Alle Rechte, auch die des Nachdrucks, der Vervielfältigung oder der Verwertung bzw. Mitteilung des Inhaltes dieses Dokumentes oder von Teilen daraus behalten wir uns vor. Kein Teil darf ohne schriftliche Genehmigung der I.R.I.S. AG in irgendeiner Form reproduziert, an Dritte weitergegeben – insbesondere unter Verwendung elektronischer Systeme – verarbeitet, vervielfältigt, verbreitet oder zu öffentlichen Wiedergaben benutzt werden.

Alle Rechte für alle Länder vorbehalten. I.R.I.S., I.R.I.S.-Produktnamen, I.R.I.S.-Logos und I.R.I.S.-Produktlogos sind Marken von I.R.I.S. Alle anderen aufgeführten Produkte und Namen sind Marken oder eingetragene Marken der entsprechenden Inhaber.