# Document Processing service (DocProc)

Part 1: Application case description

The project involves the development of a business-to-business (B2B) system for document processing. This document describes the project and the application domain, and outlines the existing technological and business constraints.

Section 1 provides some background about the domain and introduces the envisioned service. Section 2 introduces the different stakeholders of the DocProc system. Section 3 provides a detailed account of the main requirements and constraints for the document processing system.

# 1 Context and Summary

In their day-to-day operations companies consume and produce a large number of documents. Administrating this constant document stream requires considerable labour and, as a consequence, costs money. For example, a company such as Telenet must issue a payslip to each of its employees as well as deliver invoices to its millions of subscribers on a monthly basis. But also for smaller companies this document administration can be a hassle. They might have to hire extra people in their human resources department just to master the entire administrative processes, further increasing the cost of their day-to-day operations. Two trends can be seen concerning how companies try to minimise the labour and cost of their document administration.

First, there is a trend to (partially) outsource document generation, delivery and management. For example, by outsourcing human resources management to a social secretary such as Securex<sup>1</sup>, this social secretary will then take care of gathering the required data and both generating and delivering the corresponding documents based on that data.

Second, document delivery is increasingly digitised. For example, Telenet allows its subscribers to receive invoices via email, eliminating the complexity and delay of collaborating with postal services for invoice delivery. Furthermore, third party services such as Zoomit<sup>2</sup> and Doccle<sup>3</sup> have been created, allowing companies to partially outsource document delivery and monitoring the processing of these documents, e.g. keeping track of the payment in case of invoices. These services are however limited in that they only allow to outsource the actual document delivery, i.e. the last step in document processing. Moreover these services only employ a single distribution channel, their own service platform, through which they can only reach registered users. The benefits for a company are limited if the majority of their customers is not subscribed to these services. In practice, the document processing overhead might even increase since the company must now manage an additional output channel. It is clear that due to the increase of possible channels, the digitisation of document delivery has further complicated document management, especially document delivery.

# Positioning

A group of computer science graduates saw a business opportunity to take both these trends a step further. They launched a start-up, called DocProc, that aims to provide services related to automated processing of documents. DocProc will not only take care of the delivery of documents, as the above services do, but also handle their generation, archival/storage and follow-up, thus allowing companies to outsource their entire document processing flow. Figure 1 illustrates the ambition of DocProc to take a central position between companies and the recipients of their documents. From the point of view of customer organisations, their document administration will as such be drastically simplified, since they must only provide raw data to DocProc and DocProc will take care of the rest. However, as companies still want to be able to monitor their document flow, the DocProc system must provide a management dashboard to representatives of the customer organisations, and perhaps be open to further integration in the IT infrastructure of the customer organisation.

<sup>1</sup>https://www.securex.be

<sup>&</sup>lt;sup>2</sup>https://www.zoomit.be

<sup>3</sup>https://www.doccle.be

Recipients can still receive their documents via traditional channels, e.g. postal mail, email and Zoomit, which obviously must be supported by DocProc. But DocProc also provides an added value to recipients by offering an additional service to recipients who register. Registered users will have access to a personal document store (PDS) that archives all their documents processed by DocProc.

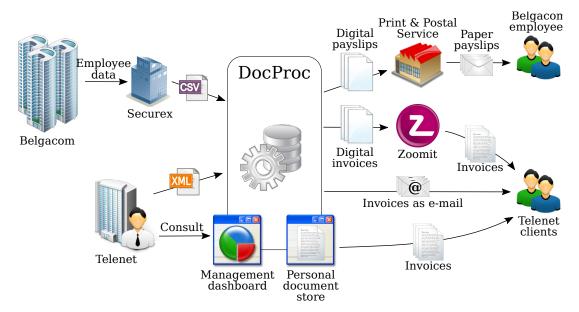


Figure 1: Overall positioning of DocProc.

The DocProc services are created in an incremental fashion. Initially DocProc focuses on invoices and payslips since these documents provide the largest immediate business opportunities.

You are contracted to create a comprehensive software architecture for DocProc. The goal is to offer the DocProc system as a service to companies. Instead of developing and selling the software as-is (i.e. per license), DocProc will set up, develop, maintain and operate its own infrastructure. As a result, DocProc will be located outside of the existing IT infrastructure of its customer organisations, who interact with the system using a web user interface (for human users) or a web service (for integration with existing IT systems).

# 2 Main stakeholders

Customer organisation The main stake of the customer organisation is that it wants to decrease the cost of its administration by outsourcing the processing of documents. Furthermore, it wants to offer an added value to their own customers by allowing a wide range of options to receive documents. A customer organisation wants to be able to easily consult the status concerning its previously processed documents and follow-up on documents being processed. Furthermore, they want to be notified as soon as possible in case of any problems. Since DocProc handles confidential data, e.g. salary information in payslips, security and privacy are also important (cf. Section 3.4).

**DocProc administrator** This is a DocProc employee responsible for managing the registration and unregistration of customer organisations. This includes conducting contract negotiations with customer organisations, providing an initial document template to customer organisations and verifying the correctness of uploaded customer organisation templates.

External email providers email providers want to deliver emails addressed to their users. DocProc relies on them to reliably deliver documents via email.

Human resources representative An employee of the customer organisation working at the human resources department. This representative is responsible for all interactions with the DocProc system concerning payslips and wants an easy and secure way to send batches of raw payslip data to the DocProc system. Furthermore, he or she requires the ability to consult the status of all requested payslips and has to be contacted by the DocProc system in case errors occur while processing payslips.

**Print & postal service** DocProc initially collaborates with a single print & postal service to both print documents and deliver them to their intended recipients. A service level agreement (SLA) between DocProc and the print & postal service determines the exact details of the provided services. This SLA determines, among others, the fee per (type of) document, the minimum number of documents the print & postal service should be able to handle, and the maximum delivery time allowed for documents, e.g. 24 hours.

**Recipients** Recipients want to receive documents addressed to them correctly and easily.

Recipients with a personal document store The DocProc system allows registered recipients to receive and consult all documents addressed to them through a web portal. These recipients want easy and secure access to their documents while avoiding the hassle of storing and archiving these documents themselves.

Sales representative An employee working at the sales department of the customer organisation. He or she is responsible for managing invoices and wants an easy and secure way to interact with the DocProc system. These interactions include sending raw invoice data and consulting the status of invoices. Furthermore, this representative is responsible for correcting any errors that occur while processing invoices.

**Social secretary** The social secretary wants to simplify the payroll administration of their clients, i.e. by managing all data required for payslips. Therefore, it wants an easy and secure manner to provide the raw data concerning payslips to the DocProc system.

**Telecom operator** Telecom operators provide a means for external users to communicate with the DocProc system and vice versa. A service level agreement between DocProc and the telecom operators determines the capabilities of these communication channels, e.g. guaranteed speed and uptime.

**Zoomit** Existing document processing services receive documents for users subscribed to their services. They are responsible for delivering the documents to the corresponding recipients and following up any further processing as required by a document, e.g. the actual payment of an invoice.

# 3 Details of DocProc

#### 3.1 Documents

As mentioned, DocProc will initially focus on invoices and payslips. The processing and intrinsicalities of these two document types are described in further detail in the following two sections.

#### 3.1.1 Invoices

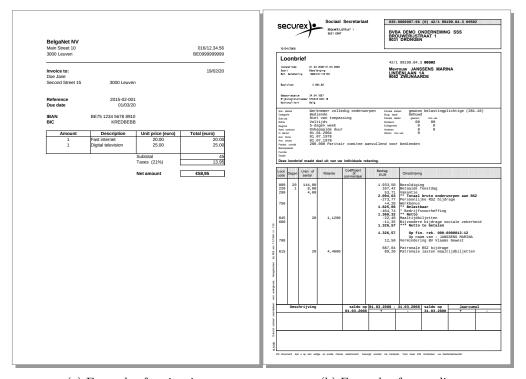
An invoice is a commercial document issued for the sale of products and/or services from a seller to a buyer. Figure 2a shows an example of an invoice. In the context of DocProc, the sellers are the DocProc customer organisations whereas the buyers are the document recipients. For example, Telenet sends an invoice to each of its customers each month (the exact day of the month depending on when the customer subscribed). To be compliant with Belgian law, invoices must contain a number of fields. An invoice must explicitly contain the term "invoice" and the date on which it was issued. The reference number of the invoice, usually formatted as a combination of the current year and a serial number, must be mentioned. Obviously, both the seller and the buyer have to be correctly identified on the invoice.

Sellers are identified by their full name, either the company name or the name of a person, the address of its head office (maatschappelijke of administratieve zetel), VAT identification number (BTW- of ondernemingsnummer), bank account number (preferably in International Bank Account Number or IBAN format) and, optionally, a Bank Identification Code (BIC). Buyers are identified by their full name, address and, if applicable, VAT identification number.

An invoice must also clearly describe and quantify the delivered goods and/or services. An invoice must mention either the date on which the goods and/or services were delivered or the date on which they are to be (partially) paid. In case an advance payment was made, the invoice must refer to the corresponding advance invoice. Every invoice must contain the gross amount (brutobedrag) to be paid in euro. This gross amount is the sum of the net amount (nettobedrag) and taxes. Each applicable tax rate must be mentioned

both as percentage and absolute amount in euro. In case of any tax exemptions, the invoice must reference the relevant laws. If there are any *minutiae* or deviations, e.g. the invoice concerns a test sale, they must be clearly described. Finally, the terms & conditions (*algemene voorwaarden*) of the seller must be appended.

Invoices are typically sent around the time the corresponding product or service is sold or delivered. Furthermore in some cases advance invoices are issued to partially pay the products and/or services up front. The timing and amount of invoices can be very erratic. For example, as indicated earlier, Telenet does not send all its invoices on the same day of the month but sends the invoice to each customer on the day of the month on which he or she subscribed. Furthermore, there can be periodic variability in number of invoices. For example, a retailer will typically see a strong increase in number of purchases, and thus invoices, during the Christmas holiday peak.



- (a) Example of an invoice.
- (b) Example of a payslip.

Figure 2: Example documents processed by DocProc.

#### 3.1.2 Payslips

Besides invoices, DocProc also handles payslips. Employers are legally obliged to provide their employees with payslips so that they can verify exactly how their salary for a certain period is calculated. Most large companies do not administer their own payroll but rely on a social secretary, e.g. Securex, for this. In such cases, it is the social secretary that gathers and combines the required data instead of the company itself and that initiates payslip generation and delivery.

Figure 2b shows an example of a payslip. According to Belgian legislation, a payslip must contain the full name of the employer and the address of its head office. The employee must be identified at least by his or her last name and initial letter of his or her first name combined with the registration number of the employee with the employer. The period to which the payslip applies must be indicated. The amount of work the employee performed within the relevant period has to be expressed in function of, for example, hours, days or products. The net salary (nettoloon) is calculated as the sum of the taxable salary and the amount exempt from taxes minus the advance tax payment (bedrijfsvoorheffing). The taxable salary is the gross salary (brutoloon) minus withholding for social security, supplemented with any amounts exempt from social security, e.g. private use of a company car. The gross salary is the salary earned based on the performed work, i.e. amount of work performed times the basic wage (basisloon), supplemented with compensation for overtime, paid holidays and any other (non-exceptional) premiums. The basic wage must be expressed using the same unit as the amount of work performed, e.g. as an hourly or monthly wage. Any amounts exempt from taxes must be listed. Finally, any reimbursements for expenses the employee has advanced for the

employer, e.g. payment of a business lunch, must be listed as well.

In most cases, employees receive a payslip once a month, usually at the at the end of each month. There are, however, exceptions concerning the recurrent behaviour of payslips. For (short) interim assignments, payslips can cover different periods of time and thus can be issued more irregularly. Furthermore, addenda to already issued payslips can be required. For example, if an employee takes sick days near the end of the month his or her (already-sent) payslip will not correctly show these. In such cases an addendum will be issued in the first half of the following month to correct the payslip sent earlier.

## 3.2 Document processing

Processing documents involves four aspects: generating, delivering, archiving, and following up on documents. Figure 3 shows how these aspects are related to each other. Document generation is the core business of DocProc and entails generating PDFs out of the raw data provided by their customer organisations. Generated documents are delivered to their addressees, which can be done through different channels. Finally, DocProc also archives all documents it produces to be consulted by its customer organisations and the recipients provided the recipient is registered with DocProc.

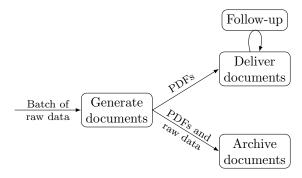


Figure 3: Illustration of the process for handling incoming raw data.

# 3.2.1 Providing the raw data

Customer organisations provide the DocProc system with the raw data required to generate and deliver the documents in question. Providing the raw data separately for each document would be cumbersome for the customer organisations as well as the DocProc system. Therefore, customer organisations deliver the necessary raw data in bulk, where batches contain the raw data for many documents. These batches are typically large. Recall the earlier Telenet example, a company that each month sends roughly 1.5 million invoices to its subscribers and approximately 2200 payslips to its employees. In rare cases, batches contain few or only a single entry, e.g. an addendum for an employee that was sick at the end of the month. A batch of raw data also includes meta data for each document. This meta data includes the desired delivery method and the corresponding required information. For example, the postal address of the recipient is provided for a document that must be delivered via postal mail. Table 1 shows an excerpt of a batch of raw data for invoices, where the last two columns contain the meta data concerning delivery.

DocProc wants to make the whole process as simple as possible for their customer organisations. Therefore, the system must indicate any errors, e.g. missing or incorrectly formatted data, as soon as possible to the customer organisations. The processing of documents for which the raw data is correct should not be influenced by errors in other documents in the same batch. Thus the customer organisation must solely correct and resubmit the raw data concerning failed documents to correct encountered errors.

Obviously different companies use different software packages for managing the relevant data, e.g. an SME (small and mid-size enterprise) might use Excel whereas a large multinational often has specialised software. Therefore, Docproc must accept a wide variety of input formats, e.g. CSV, XML and Excel files. Customer organisations can send batches of raw data either manually via their management dashboard (cf. Section 3.3) or allow their own system to communicate with the Docproc system via several protocols (e.g. SCP or FTP) and further integrate this in their own enterprise work flows. Note that each access to the Docproc system must be authenticated (cf. Section 3.4).

Reference	Issue date	Due date	Last name	First name	 Delivery	Delivery data
2021-02-1	19/02/21	01/03/21	Doe	Jane	 Postal mail	Jane Doe Second Street 15 3000 Leuven
2021-02-2	19/02/21	25/02/21	Smith	John	 Zoomit	BE61310126985517
:	:	:	:	:	 :	:
2021-02-45	23/02/21	10/03/21	Anders	Jef	 email	jef.anders@email.com
:	:	:	:	:	 :	i i

Table 1: Excerpt of a batch of raw data for invoices.

#### 3.2.2 Generating the documents

After receiving the raw data, the DocProc system generates the actual documents from this data. In order to allow customer organisations to give its documents a distinctive look and feel they provide their own template. A template is provided as a Microsoft Word Open XML Document (docx) file, which can be instantiated with raw data. An excerpt of a simple template for invoices is shown in Figure 4. In this template, "parameter" indicates all values to be filled in when instantiating the template. Note that all static data, e.g. company name, is already filled in the template. The output of the document generation step is a PDF file which is subsequently sent to its intended recipient and archived by DocProc. The exact generation process is different for invoices and payslips, as explained in the following sections.



Figure 4: Excerpt of a customer organisation template for an invoice.

**Invoices** The process of generating an invoice is illustrated in Figure 5. In the first step the customer organisation template is filled in using the raw data provided by the customer organisation. Any encountered errors are reported to the customer organisation so that it can address the problems. For example, the raw data can be badly formatted, e.g. a missing close-tag in an XML file, or there is missing data for completing the template, e.g. no gross amount to be paid.

Based on a completely filled-in template, a PDF version of the invoice is generated. Since invoices have legal value for the customer organisations they want to make sure nobody can tamper with it after sending the invoices. Therefore, each generated PDF is signed with the key of the customer organisation. The signed PDF file is sent to its intended recipient using the delivery meta data provided as part of the raw data. Finally, the DocProc system also archives the raw data along with the invoice generated from the data. To comply with Belgian legislation the DocProc stores invoices for at least seven years, although customer organisations can negotiate for a longer term.

Payslips The process of generating payslips is illustrated in Figure 6. The customer organisation template is first instantiated using the provided raw data. If any errors, e.g. missing data or the basic wage and amount of performed work are expressed in different units, are encountered they are reported to the customer organisation in order to be resolved.

After correctly filling in the template with raw data, a PDF version is generated for the payslip. The generated PDF file is, based on the delivery meta data, consequently sent to its intended recipient. Finally,

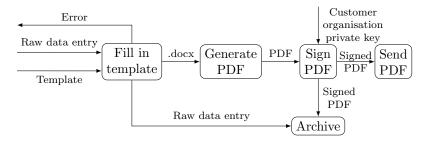


Figure 5: Illustration of the invoice generation process, indicating the different steps and the data they require.

DOCPROC archives the raw data and PDF file generated from this data.

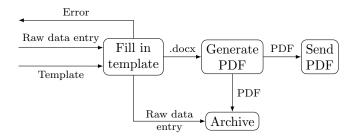


Figure 6: Illustration of the payslip generation process, indicating the different steps and the data they require.

#### 3.2.3 Document storage

As mentioned DocProc stores the generated documents along with the raw data they were generated from. Customer organisations can use these stored documents in their overall workflow, e.g. to compose revenue or tax reports. The customer organisations access these documents through their management dashboard (cf. Section 3.3).

Personal document store As an additional service, DocProc allows recipients to register with the DocProc system and provides them with a personal document store (PDS). The personal document store also contains all documents previously delivered to the email address the recipient registers with. Furthermore, the personal document store allows to search for specific documents, e.g. all invoices from a certain store or received at a particular date, and download the PDF files if desired. Recipients might consult these to, for example, fill out their yearly tax forms or to keep track of their monthly spending.

# 3.2.4 Delivering the documents

Generated documents have to be delivered to their intended recipients. In order to become successful, DocProc must be able to reach a critical mass of recipients and thus the DocProc system has to be open concerning the supported delivery methods. The customer organisation indicates the desired method for each document, including the information required for this method, as meta data in the raw data (cf. the last two columns of Table 1). For example, if a document should be delivered via email the customer organisation must provide the email address of the recipient. The customer organisation must keep track of how each of their recipients want to receive his or her documents.

Receipt tracking As an added service to their customer organisations DocProc allows to enable receipt tracking of invoices for those delivery methods that support it (e.g. email). If receipt tracking is enabled, customer organisations can consult the status through the management dashboard (cf. Section 3.3). How the receipt tracking is implemented depends on the selected delivery method.

The following delivery options should be supported in a first version of the DocProc system.

**Print & postal service** Delivering documents on physical paper via a print & postal service is the first, more traditional option. In this case, the costumer organisation has to provide the postal address as meta data. DocProc sends PDFs including the complete postal address of the recipient to the print & postal service, via a web service, which in turn prints, packages the documents and delivers them to the recipient via postal mail. While receipt tracking is supported by postal mail using registered mail (aangetekende zending), DocProc currently does not want to provide this yet.

Email More environmentally-conscious customer organisations can allow their recipients to receive documents via email. In this case, the recipient must share their email address with the customer organisation and the customer organisation provides the email address in the meta data. Depending on whether receipt tracking is enabled, two alternative ways to deliver documents via email should be supported. Without receipt tracking, the email in question will contain a short message indicating that there is a new document from the sender and the document itself as attachment. With receipt tracking, the email contains a short description of the document (type, sender, etc.) and a unique URL locating the actual document in the DocProc system. When the recipient follows the URL to retrieve the document the DocProc system knows it is delivered and can inform the customer organisation. A unique URL is only valid for 30 days. If a unique link expires before the recipient retrieves his or her document this is indicated in the management dashboard (cf. Section 3.3).

**Personal document store** From the point of view of a recipient the personal document store replaces email as document delivery method. Instead of delivering the document via email to the recipient, it is added to his or her personal document store and a notification is sent to him or her concerning the availability of the new document. Since the personal document store is located within the DocProc system, receipt tracking is provided for registered users to the customer organisations.

Zoomit One of the main external service providers DocProc collaborates with is Zoomit. Zoomit is a free service for Internet and Mobile banking which allows its users to manage invoices more easily. Companies can register with Zoomit to send invoices to those recipients that activated Zoomit. Interested users can activate Zoomit through their PC banking application, after which their bank account number serves as their identifier. After activation, users can select from which Zoomit partners, referred to as senders, they want to receive invoices via Zoomit. Since Zoomit uses the bank account number as identifier, users must have performed at least one payment from this account to the sender before they can request to receive their invoices via Zoomit. Companies are informed by Zoomit when one of their recipients has selected it as allowed sender and receive the identifier of the recipient, i.e. the name and bank account number. The company will from now on deliver invoices for the recipient via Zoomit. If a user no longer wishes to use Zoomit for a specific sender they can withdraw this sender's permission at any time. Zoomit will inform the sender that the user is no longer interested in receiving his or her invoices via Zoomit. In that case, the sender and user must agree upon a new way to send invoices. Zoomit keeps track of the status of each received document and allows the sender to query this information.

# 3.3 Management dashboard

Although customer organisations outsource the actual processing of their documents they still want to be able to consult them and follow up on the process. Therefore, DocProc provides a management dashboard through which the appropriate employees of the customer organisation, such as billing responsibles, can consult the status of their documents. The status of each document can range from "raw data received" to "sent to recipient". If any errors were encountered, e.g. the raw data for a document was incomplete, they are indicated. Based on these error reports the customer organisation can take appropriate actions, e.g. resubmit a new batch containing corrected data.

If the customer organisation has enabled receipt tracking, the dashboard will also indicate whether or not the recipient has effectively received the document. Note that receipt tracking is only supported for email, Zoomit or the DocProc personal document store (cf. Section 3.2.4).

In addition, the dashboard offers additional possibilities, such as consulting and updating document template(s) and consulting billing information.

#### 3.4 Additional constraints

In addition to the general problem domain description, there are a number of specific constraints highly relevant to DocProc.

#### 3.4.1 Service Level Agreements with customer organisations

With each customer organisation DocProc negotiates a service level agreement (SLA), stipulating the exact services provided to the customer organisation. For recurring batches of documents this SLA determines the fixed periodical deadline. For example, the payslips of the customer organisation must be processed on the 28th of each month. To ensure the periodical deadline is achieved the customer organisation agrees to provide all required data at least 24 hours before the deadline. Furthermore, the SLA will state the fixed price and the maximal size, i.e. number of documents, for each recurring batch.

For non-recurring batches the SLA contains the default priority. To meet the different needs of customer organisations, DocProc offers three priorities: diamond, gold and silver. These priorities correspond to generation time of maximum 12, 24 or 48 hours respectively. If necessary, customer organisations can change the priority of individual documents to a different priority. In case of emergencies, e.g. an overdue invoice that must be sent as soon as possible, DocProc offers the possibility to mark a document as critical. Such critical documents are to be processed within a 5 hour deadline. The price for each document depends on the chosen priority.

Furthermore, the SLA stipulates quality requirements concerning availability. Obviously no raw data or documents may be lost anywhere while processing documents, including documents exchanged with third parties such as Zoomit.

Finally, the SLA between DocProc and a customer organisation stipulates some other agreements. For example, DocProc will not be liable in cases where deadlines are missed due to incorrect or incomplete raw data delivered by the customer organisation. Furthermore, the SLA contains the required parameters for documents which should be printed, i.e. type of paper, single- or double-sided and colour or black-and-white.

#### 3.4.2 Security and Privacy

Since the data and documents processed by DocProc concerns both financial information of both customer organisations and recipients and personal data of end users (*data subjects*), it must be appropriately protected. In other words, security is a key concern of the DocProc system. In a first instance, the scope is limited to basic *user authorisation* and *user authentication* at the management dashboard (cf. Section 3.3) and the personal document store (cf. Section 3.2.4).