Leveraging corporate context within knowledge-based document analysis and understanding

CLAUDIA WENZEL, HEIKO MAUS
German Research Center for Artificial Intelligence (DFKI)
P.O. Box 2080, 67608 Kaiserslautern/Germany
E-mail: wenzel@dfki.uni-kl.de, maus@dfki.uni-kl.de

Abstract

Knowledge-based systems for document analysis and understanding are quite useful whenever analysis has to deal with varying free-form document types which require distinct analysis components. In such a case, declarative modeling is a good way to achieve flexibility. An important application domain for such systems is the business letter domain. Here, high accuracy and the correct assignment to the right people and the right processes is a crucial success factor.
Our solution to that proposes a comprehensive knowledge-centered approach: We model not only comparatively static knowledge concerning document properties and analysis results within the same declarative formalism, but we also include the analysis task and the current context of the system environment within the same formalism. This allows an easy definition of new analysis tasks and also an efficient and accurate analysis by using expectations about incoming documents as context information. The approach described has been implemented within the VOPR[1] system. This DAU system gains the required context information from a commercial workflow management system (WfMS) by constant exchanges of expectations and analysis tasks. Further interaction between these two systems covers the delivery of results from DAU to the WfMS and the delivery of corrected results vice versa.

Keywords: document analysis system, office processes, context information, document knowledge

1 Introduction

The analysis of printed business letters has become a lucrative application domain for document analysis and understanding (DAU) techniques within the last few years. Meanwhile, some commercial systems are available on the market for this purpose, e.g., AscentCapture by Kofax [1], TELEform from Cardiff [2], or Free Form from Captiva [3]. The goal of such systems is to analyse incoming business letters in order to extract all relevant information and to assign them to the right people working within the right processes. This saves inhouse mail distribution and manual indexing efforts. Another commercial approach which classifies OCR documents and extracts the relevant information, is SERbrainware by SER [4]. Its focus is a little bit different, since this system aims at incorporating knowledge from paper documents into a company‘s knowledge management system. However, the underlying techniques are rather similar.

In order to adapt DAU systems to changing and specific requirements of distinct companies, most of these commercial systems (as far as details are published) keep the layout, logical, and textual knowledge about the documents within the current company‘s domain in a declarative style. We refer to such systems as knowledge-based document analysis and understanding (DAU) systems.

Besides document properties, there is other knowledge available which is useful for analysis and can be kept declarative such as the analysis task or the corporate context. We argue that declarative modeling as far as possible leads to several advantages which are especially useful in

1. VOPR is an acronym for the Virtual Office PRototype.

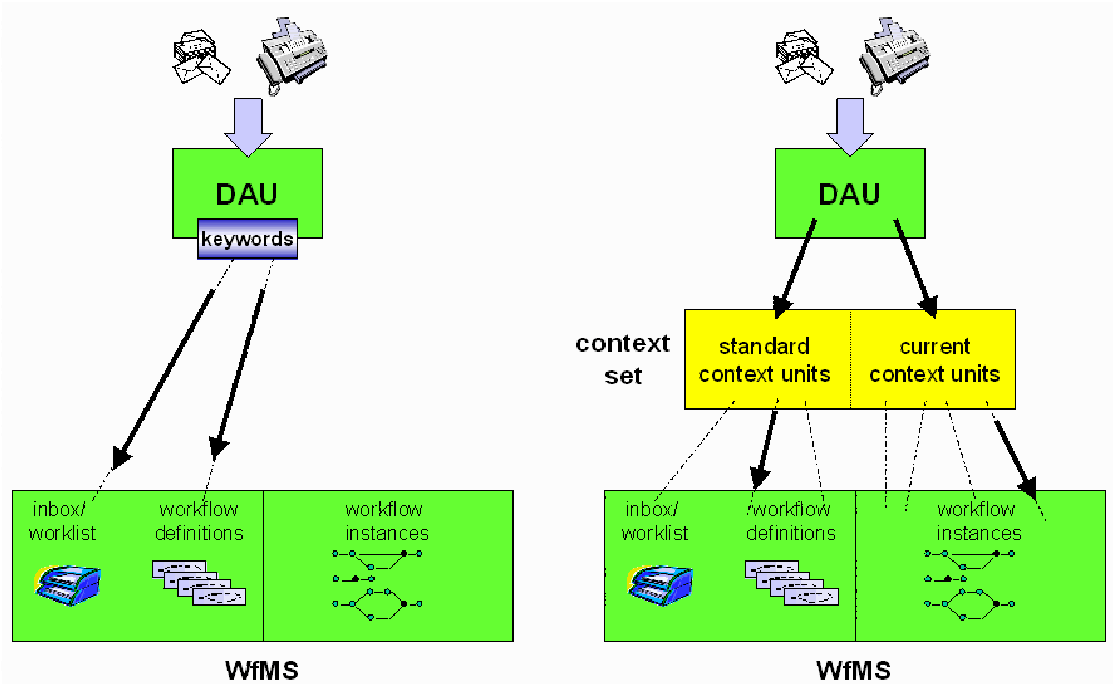


Figure 1: State-of-the-art procedure in the analysis of printed business letters (left-hand side) in comparison to the VOPR approach (right-hand side)

real world applications.

To this end, figure 1 explains the typical way of document analysis for printed business letters on the left-hand side and compares it to our own VOPR approach. Each way starts with incoming mail which is fed to a system for document analysis and understanding. Then, the DAU system analyses the documents according to a specific analysis task. The specification of this analysis task already reveals a difference between the two views: In the traditional view, the analysis task is static and predefined to a special extent: For the goal of process assignment, some keywords have been connected to the names of employees within the company. Such relations might be „Invoices -> Mrs. Snider“ (which means that Mrs. Snider treats all invoices) or „Company Microsoft -> Mr. Gates“ (which means that Mr. Gates treats all letters coming from the company Microsoft). The kinds of keywords to be used are predefined (e.g., keywords must be company names) and are system-internally hardwired to the analysis task (e.g., sender extraction) and to analysis components (e.g., a pattern matcher). Thus, some specific analysis tasks are employed to determine information items which relate the document in a static way to people within the company. Moreover, a message frame has been attached to each keyword which specifies the kind of information to be extracted from the document for the employee. The message frame itself is also directly related to a specific analysis component which is able to process the frame‘s syntax.

Thus, the overall analysis is accomplished in two steps: In the first step, the matching keyword (and therefore the corresponding employee and the message frame) is determined. In the second step, all information denoted in the message frame is extracted from the letter. Finally, the document along with its filled message frame is sent to the corresponding employee. To sum up, a predefined kind of information given by the keyword is used as the only connection between the incoming letter and the system environment within the company.

However, the system environment is more than just a person‘s inbox. In today‘s companies, administrative procedures dealing with business letters are often automatized with the aid of workflow management systems (WfMS for short). Such WfMS have the task to get the right data to the right people with the right tools at the right time [5]. To this end, they support employees within their daily work by automating their business processes as workflow instances. Thus, instead of assigning incoming mail always to a mail tool of specific persons, it makes sense to directly attach incoming mail to the corresponding workflow instances. In this case, people involved get the data directly with the right tool at the right place. Such a solution is subject of this paper and shown at the right-hand side of figure 1.

Roughly speaking, a *context set* is steadily collected by looking at all open workflow instances (current context units) and by modeling default cases (standard context units). This context set represents all currently possible analysis tasks as well as application context from the workflow instances. The application context states the expectations of the WfMS to future incoming documents such as the document‘s type, its sender, or its topics. The analysis of incoming letters uses the context set for a goal-directed analysis and attaches letters along with extracted information usually directly to waiting workflow instances. Just in case that an incoming document starts a new process (e.g., a request) or is only related to certain persons, the system attaches letters to people‘s inboxes or to default workflow definitions.

As we will prove later-on, this VOPR approach has some crucial advantages:

- The declarative modeling of application context, analysis task and message frames enables an easy adaptation to new kinds of documents and new message frames, but also to new company processes.
- The semantic application context can be used to restrict the search space for analysis components. This allows for a higher analysis efficiency.
- For the same reason, we also reach a higher accuracy.

The remainder of this paper describes the VOPR system in more detail. We start with a description of related work in the next chapter. Afterwards, chapter 3 explains the underlying system architecture. Then, all main components are explained: Chapter 4 explains the context collection within a WfMS and its delivery to the DAU system. The integration of context within the knowledge base of the DAU system is subject to chapter 5, while chapter 6 discusses the expectation-based execution of analysis algorithms by our analysis control. Chapter 7 details the analysis algorithms employed and chapter 8 explains their relation to different analysis strategies. The paper is completed by some experimental analysis results in chapter 9 and some concluding remarks in chapter 10.

2 Related work

Within document analysis literature, the usage of application context is not very common. Of course, a lot of systems use context coded as heuristics directly within the analysis procedure, but this allows only a strict analysis and there are no means for exchanging such heuristics. The only context usage described is based on data available within databases. Such data is typically used for the verification of OCR results, e.g., legal account numbers, customer names and bank codes within bank cheques.

However, there are commercial as well as research systems which deal with the analysis of business letters. One commercial system for which results have been published is described by Bleisinger et al. [6]. Their IntelliDoc system combines a declarative description of document properties with context from databases. It has been developed for the analysis of more or less structured business letters. Document properties concerning layout, logic, and content are entered within a graphical user interface by using a system-specific syntax. Within the same interface, connections to valid entries in databases can be established. Unfortunately, details about this integration are not mentioned. During analysis, document knowledge is used by a speech recognizing component and a syntactic pattern-matcher.

Another research approach which is meanwhile exploited commercially is the FRESCO formalism (frame representation of structured documents) developed by Thomas Bayer [7]. It has been applied to German and English business letters. It allows the definition of knowledge about document components as well as defining knowledge about analysis algorithms. Documents and their parts along with layout and logical properties are represented by concepts (for generic documents resp. parts) and instances (for concrete images). Analysis components are described in a taxonomical hierarchy by defining application properties. During analysis, a global scheduler chooses analysis components. These components generate instances for document parts and fuzzy-based inference mechanisms combine instances to valid document instances.

In a similar way, our own previous work within the Omega project, see e.g., [8] has been put on the market. Our research prototype contained a toolbox of task-specific analysis specialists for image preprocessing of gray-scale documents, OCR, voting, morphological analysis, logical labeling, document classification and so on. These specialists communicated with the aid of a blackboard architecture and worked on the domain of German business letters. The Omega concept has been the basis for the development of the SmartFix product [9] which today serves several markets, e.g., insurance accounts,

Also, a system described by Bläsius et al. [10] has been commercialized within the WANDO product [11] for the analysis of invoices. From a scientific point of view, it differs from the FRESCO approach mainly by the usage of the dempster-shafer approach instead of fuzzy sets for the propagation of uncertainty values.

Other research which might be listed within the context of knowledge-based document analysis has been done by Stephen Lam [12] who describes a frame-based approach especially for logical labeling. Further representation formalisms for generic document analysis purposes are based on neural nets [4], predicate logic [13] and production rules [14].

3 System architecture

The full-fledged integration of a DAU system into a WfMS-based business environment is displayed in figure 2. Data flow starts on the one hand with a clerk's input to the WfMS (for the derivation of the application context and the analysis task) and on the other hand with new incoming documents (for the true analysis). New incoming documents already scanned and in an electronic representation are directly transferred to the DAU system by a post-box server.

The DAU system mainly contains components for a typical knowledge-based DAU: Central

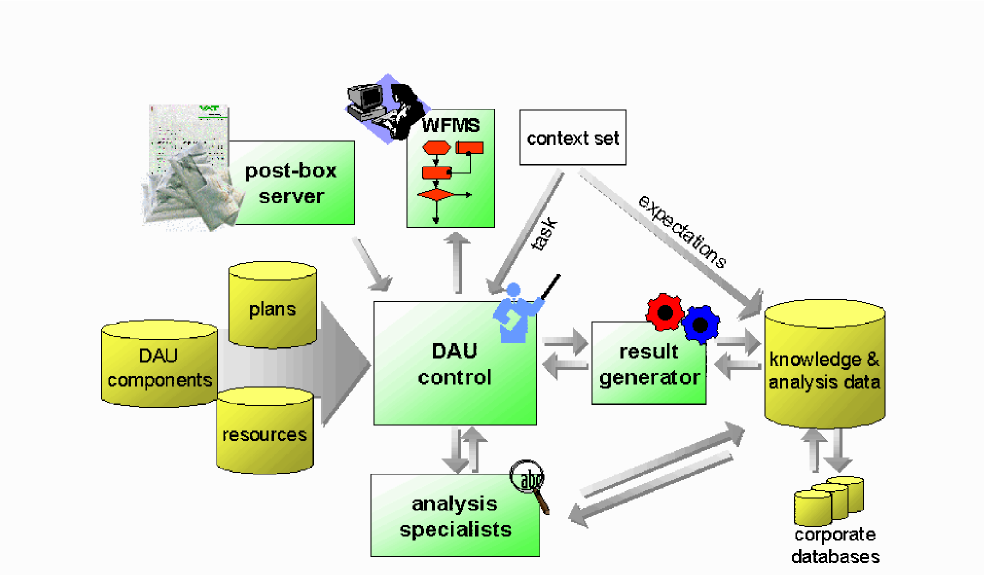


Figure 2: System architecture of the VOPR system

part is the DAU control which communicates with the post-box server and the WfMS which automates the underlying business processes. The DAU control uses resources and DAU components to assemble analysis specialists used by task-specific plans to accomplish the analysis tasks at hand. The analysis specialists use the knowledge and analysis data within the document knowledge and incorporated databases (e.g., postal addresses of suppliers) to fulfill their specific analysis tasks and to store their analysis hypotheses. These hypotheses are input to the result generator which unifies them to one final result offered to the requesting workflow. Afterwards, the analysis control returns its results to the waiting workflow instance which presents them to the user. He or she corrects wrong DAU results which may later-on lead to new document knowledge items derived by different learning components.

In order to enable the system to cope with context information, several components have been extended with appropriate interfaces. First, a WfMS is introduced and extended with a context set in order to deliver context information to the document knowledge by means of expectations. To allow a consistent semantical interpretation of context information, a domain ontology is established which describes the structure and contents of all documents. In addition, the DAU components are able to use these expectations for fulfilling their tasks more efficiently and more accurately (e.g., by search space reduction). And last but not least, declarative analysis tasks are introduced to cope with the requirements of a dynamic environment.

The WfMS supplies the DAU system with context from workflow instances. This kind of corporate knowledge is annotated by entries of a separate corporate database. Thus, the DAU system receives relevant expectations about expected documents from the WfMS and enriches them by retrieving information from the corporate database. When the DAU system is invoked, incoming documents are analysed by different components explicitly selected by the analysis control according to the current task.

Within the following chapters, we explain this procedure in more detail.

4 Retrieving context from workflows

The retrieval of context information from workflows for DAU purposes demands two prerequisites. First, an integration of the VOPR system into WfMS has to be established. Moreover, the necessary information pieces have to be modeled. These topics are subject to the following paragraphs *WfMS integration* and *Information representation*. Given these prerequisites, the retrieval of context information divides into two major steps, namely *context collection within workflows* and *context transformation into context units* - a representation suitable for the DAU‘s knowledge base. These steps are shown in figure 3 and detailed in the second part of this chapter. The description of a specific case of context units - the so-called *standard context units* - ends this chapter.

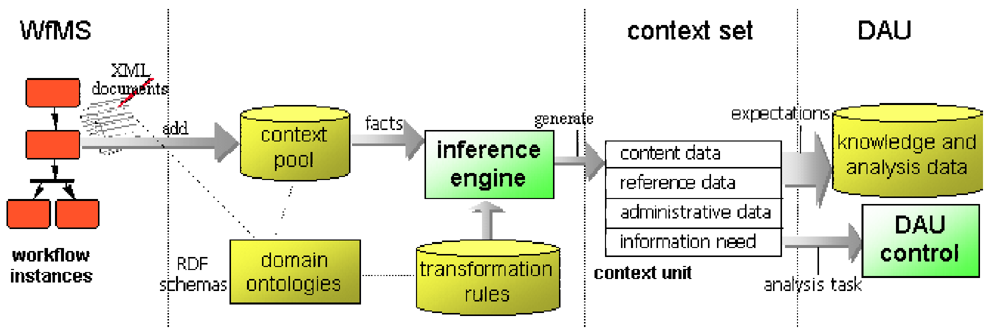


Figure 3: Generation of context information from WfMS

WfMS integration. Since today‘s (commercial) WfMS do not have modeling concepts for workflow context [15], there is no possibility to directly access context as a whole at runtime. However, there are several sources, e.g., control flow, data flow, application data, or a workflow‘s history [5], which allow the collection of context information for further processing.

Because the VOPR system is supposed to implement a generic integration into WfMS independent of any proprietary extension to the workflow model, we introduce a database to which we refer as *context pool*. Therein, all valuable context information is collected throughout the workflow execution. Technically, we model special activities in the workflow definition which store involved information in the context pool, thus guaranteeing the availability of context when needed (for more details on the workflow integration concept please refer to [16]).

Information representation. In order to describe the documents produced according to our business letter ontology we need a representation for documents in the workflow. Therefore, we need a description formalism for the documents themselves, a description of the corresponding ontology, and connections of the documents contents‘ to this ontology.

To this end, we use the World Wide Web Consortium‘s (W3C) standard XML (eXtensible Markup Language). XML is a markup language for documents such as business letters containing structured information. It allows to define domain-specific, semantic, structured markup languages [17]. One such application of XML is RDF (Resource Description Framework, [18]) which provides means to describe metadata about (web-) resources, thus about our XML documents. RDF allows for the definition of domain-specific vocabularies in so-called *RDF schemas*. These schemas can be used as ontologies for XML documents. Therefore, the business letter domain is modeled as an RDF schema and is used to semantically enrich XML documents used in the workflow instances by relating the documents‘ data to this ontology.

For instance, figure 4 shows an excerpt from an order related to our business letter RDF schema (as can be seen within the first tag `rdf:Description`). The tag `Warenbezeichnung` represents a particular good which consists of several attributes. The one shown is the article name (`Artikelname`) with type `string` and value `CP 1.0-63-05L`.

```
<Warenbezeichnung>
<rdf:Description xmlns=“http://www.dfki.de/NH/Geschaeftsbrief/Leistungsangaben/Warenangaben/
Warenbezeichnung.slots#“>
     <Artikelname>
       <rdf:Description rdf:value=“CP1.0-63-05L“>
         <rdf:type rdf:resource=“http://www.dfki.de/NH/Types#String“/>
       </rdf:Description>
     </Artikelname>
    ...
</rdf:Description>
</Warenbezeichnung>
```

Figure 4: Excerpt from an XML/RDF order

Due to the possibility to use namespaces in XML, several different vocabularies can be incorporated within a single document and used throughout the workflow. This provides a powerful means for semantics in WfMS and for intelligent assistants as proposed in [19].

Given this, we are able to identify relevant information within the workflow and store it in the context pool. Thus, the information stored within the context pool is semantically described in terms of a domain ontology.

The following paragraphs will detail the aforementioned process of context generation.

Context collection within workflows. In case a situation occurs in a workflow instance (e.g., issuing an order) which causes the future arrival of a document (e.g., an invoice), the workflow states an expectation by means of a context unit. For instance, this is accomplished by the first two steps in figure 5 showing an excerpt of an actual workflow definition used for the VOPR-system integration into the WfMS Staffware 2000 [20].To be able to state an expectation, the workflow stores any information about this event into the context pool such as the document itself. Afterwards, it invokes an inference engine and hands over some administrative data, such as the information need and the context pool reference of the causing event. Then the workflow waits until its expectation is satisfied, in our example, until the event INVOARVD is triggered.

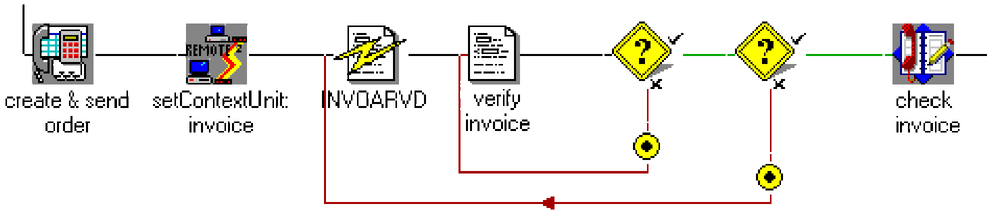


Figure 5: Excerpt from a workflow definition with VOPR system integration

Context transformation into a context unit. The concluding inference step uses the context pool as fact base for the generation of a context unit. The inference engine[1] converts the raw context information into a context unit by using transformation rules. These rules provide a mapping between the workflow‘s domain ontologies and a description of the expected document in terms of the DAU domain ontology (represented by the DAU‘s knowledge base). The three exemplary rules given in figure 6 show the ontology mapping from a fact with an id given by our business letter RDF schema into a valid path notation for the document knowledge (the path notation is <concept>(.<part>)*.**message.**<slot>).

```
(defrule  Warenbezeichnung_Artikelname
  (RDFFact    (id    “http://www.dfki.de/NH/Geschaeftsbrief/Leistungsangaben/Warenangaben/Warenbezeichnung.slots#Artikelname“)
     (type ?t) (value ?v) (package ?p)
  =>
  (bind ?path (str-cat ?*concept* “.Leistungsangaben.Warenangaben.Warenbezeichnung.message.Artikelname“))
  (addEntry ?path string (convertValue ?t ?v) ?p ?*expectationContent*))

(defrule  Absender
  (unique (RDFFact (id “http://www.dfki.de/NH/Geschaeftsbrief.slots#Empfaenger“) (type ?t) (value ?v) (package ?p)))
  =>
  (bind ?path (str-cat ?*concept* “.message.Absender“))
  (addEntry ?path integer (convertValue ?t ?v) ?p ?*expectationContent*))

(defrule Bezugsangaben_Belegdatum
  (RDFFact (id “http://www.dfki.de/NH/Geschaeftsbrief/Belegangaben.slots#Schreibdatum“) (type ?t) (value ?v) (package ?p))
  =>
  (bind ?path (str-cat ?*concept* “.Bezugsangaben.message.Datum“))
  (addEntry ?path integer (convertValue ?t ?v) ?*cpBasedUponID* ?*expectationReference*))
```

Figure 6: Simple exemplary rules

For instance, the first rule named Warenbezeichnung_Artikelname fires if a fact in the context pool has the id specified. The rule header binds type, value, and package to the respective variables where *package* denotes a number which indicates the sequence within a set-typed slot. The rule body performs several actions including the arrangement of the new concept path according to DAU‘s ontology, the conversion of the value from the RDF schema to a value of the DAU‘s ontology, and finally, the insertion of the construct into the expectation contents of the current context unit.

To derive content and meaning of the expected document, some rules must also be domain-specific, e.g., the second exemplary rule stating that the recipient (Empfänger) of the outgoing document is to be the sender (Absender) of the document expected, or the last rule assembling information from the outgoing document to describe potential references within the answer, e.g., the date when a document has been written (Schreibdatum).

This information is stored within the content and reference data part of the context unit. Other rules generate administrative data or integrate the information need of the workflow into a context unit by stating the analysis task. The resulting context unit provides all necessary

1. Actually, we use JESS (Java Expert System Shell), see [21]

information in terms of the DAU‘s knowledge representation (to be detailed in chapter 5). Figure 7 presents an exemplary context unit with paths in the German notation of the DAU‘s knowledge representation. For example, the path

Geschäftsbrief.Leistungsangaben.Warenangaben.Warenbezeichnung.message.Artikelname

translates to

businessLetter.service.goods.goodDescription.message.articleName

A context unit

- describes content and meaning of an expected document and its relationship to the business process which it refers to by stating all facts known in advance about a document. This may be the document‘s message type (1) (invoice = `Rechnung`) or the expected product list (2), because these products have been ordered in the corresponding business process. Furthermore, references to preceding documents are included, such as information from the corresponding offer (3).

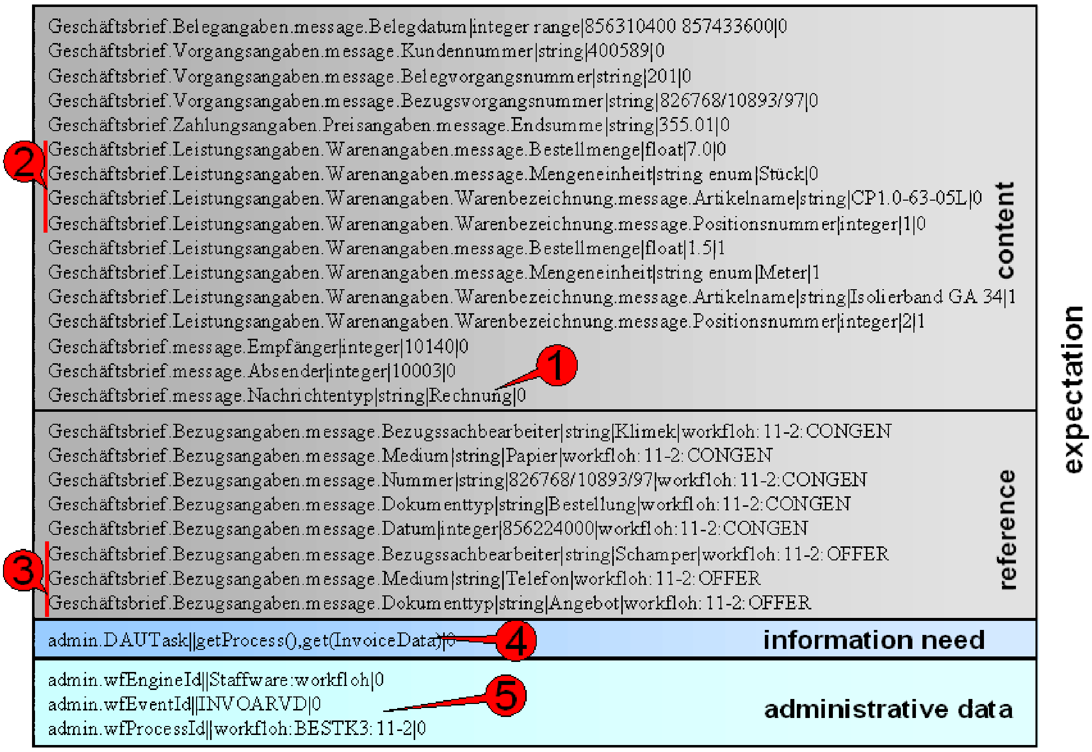


Figure 7: An exemplary context unit

- expresses the information need of a business process. This is achieved by including analysis tasks which describe the required information (4). We distinguish two kinds of tasks: The first one is called *process identification* (`getProcess()`) and is given inherently in the business letter domain because incoming documents relate to business processes to which they have to be assigned. The second one is the task of *information extraction* (`get(...)`) which requests specific information from the document necessary for further processing in the workflow. Information items required are either specified by a path notation revealing which information items have to be extracted or (as seen in the example) it is defined as a shortcut by macros which stand for a number of such paths.
- lists administrative data to accomplish the integration into WfMS such as the identification of the workflow instance and the event to trigger if the document arrives (5), e.g., in figure 5 this would be triggering the event `INVOARVD`.

The context unit generated is stored within the context set. In addition, content and reference data are inserted as an expectation into the DAU‘s knowledge representation called *document* *knowledge*. Representation and usage of an expectation in the document knowledge are explained in chapter 5.

Standard context units. Besides using dynamic context information from business processes, we also use more static context information by means of *standard context units*. These units serve two purposes: First, they describe documents not expected by a specific running business process such as adverts. Second, they serve as an exception handling for the VOPR system in two cases. If there is no match between an incoming document and context units from the WfMS, a standard context unit is the default match (e.g., for documents of type 'invoice'). Such a unit defines the default handling as, for instance, the instantiation of a workflow definition which will route an unassigned invoice to the accounting department for manual assignment. Second, standard context units catch documents referring to an existing process which are not expected such as reminders. Here, a specific analysis task has to determine the correct process in case that no explicit context unit exists for this document. Given this default handling for incoming documents, the VOPR system realizes the state-of-the-art procedure for mail processing (cf. figure 1).

5 Representing document knowledge

There is one central repository which stores generic document properties and analysis results called *document knowledge*. Besides storing typical document knowledge and data, some powerful mechanisms have been included which allow the storage of context information within the same formalism. This enables a direct context access for all analysis components and a semantic exploitation without doing any format adaptations.

Our knowledge representation formalism is a frame-based notation where frames represent documents or document parts. Possible relations between frames are specialisations (*is-a*) and aggregations (*parts*). Each frame represents an object which may be a whole document or a part of a document. It consists of slots which specify certain document properties by facets and facet values. Frames are a very well-known object-centered representation formalism in the AI community and have already been successfully applied for DAU purposes, e.g. [7].

Typical contents of document models are well-known in the document analysis community [22] [23]. Within our model, we distinguish between image (e.g., background-color), layout (e.g., coordinates), content (e.g., language), logic (e.g., syntactic text patterns) and message (all relevant semantic information) properties of a document.

There are also different types of facets which may be used for slots, namely for representing uncertainty (`:certainty`), values and value types (`:value`, `:range`, `:unit`, ...), related thesaurus information (`:thesaurus-name`, `:thesaurus-value`, ...), frequencies and relevances. Besides this, we also have some special constructs such as check-rules defining restrictions between slots and frames to ensure consistency.

The formalism is used for storing knowledge gained within a separate learning step automatically (which is not subject of this paper, but mentioned in [27]), for retrieving general properties and analysis results and - most important - for combining single analysis results for document parts to consistent overall results for a whole image under consideration.

Generic document properties are represented in frames called *concepts*. One example for a company logo is given in figure 8.This figure shows a screenshot of our document knowledge browser: In the upper right area, you can select which kinds of frames to investigate (in our case, `Concept` has been chosen). On the left-hand side, you see all frames of the selected kind which are currently defined (e.g., the first one `Papier-Belegangaben` means paper-record data). Last but not least, the lower right area contains the definition of the selected frame (`Papier-Firmen-logo` means a company logo on a paper document). The check-rules defined in this example

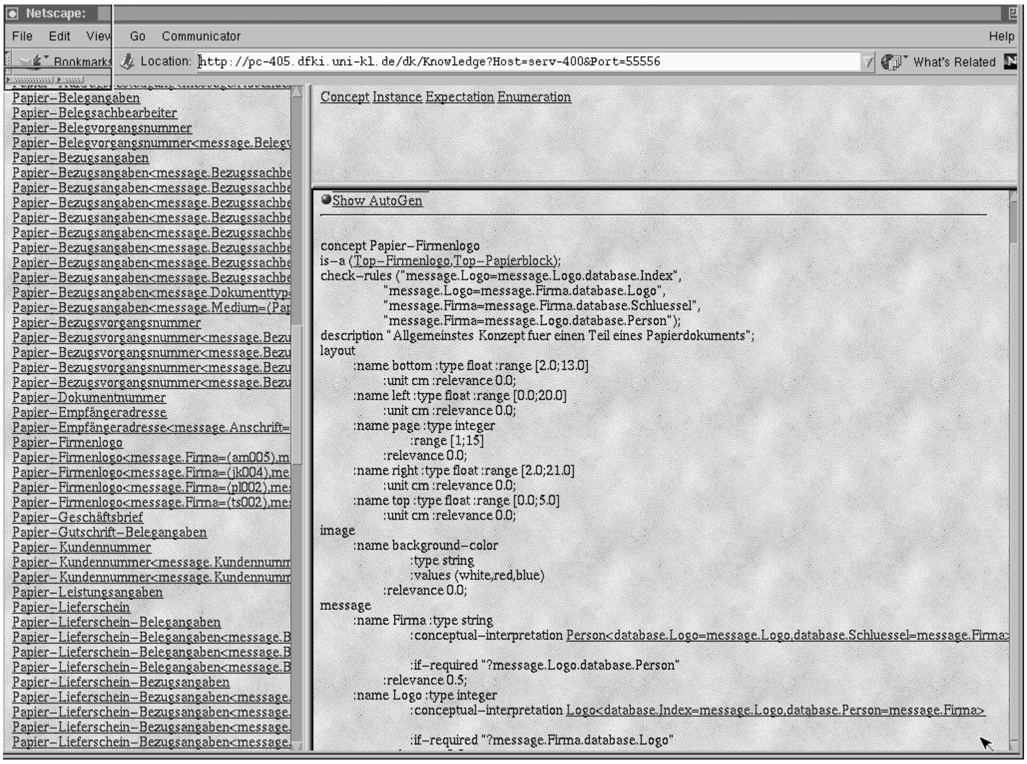
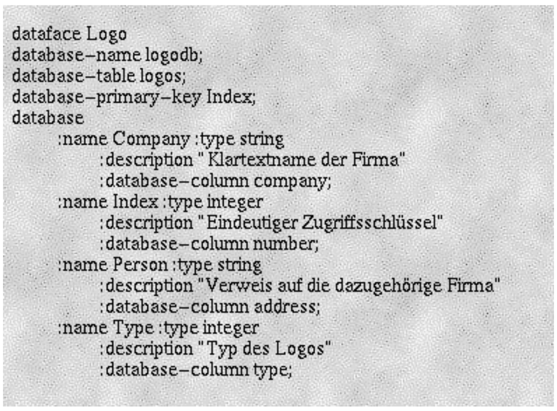


Figure 8: Screenshot of the document knowledge browser for the concept company logo

state that the company the logo belongs to must be consistent with the sender of the document.

Within the message slotgroup of the example, you see how context from databases is introduced: Within the slot `Logo`, the facet `:conceptual-interpretation` defines a link to another frame (named `Logo`) representing the structure of a database table. Such a frame type is called *dataface*. Figure 9 shows the dataface for a Logo as another screenshot of the lower right area of our browser. This dataface relates slots within the document knowledge directly to columns



```
dataface Logo
database–name logodb;
database–table logos;
database–primary–key Index;
database
      :name Company :type string
            :description "Klartextname der Firma"
            :database–column company;
      :name Index :type integer
            :description "Eindeutiger Zugriffsschlüssel"
            :database–column number;
      :name Person :type string
            :description "Verweis auf die dazugehörige Firma"
            :database–column address;
      :name Type :type integer
            :description "Typ des Logos"
            :database–column type;
```

Figure 9: Screenshot of the document knowledge browser for the dataface logo

of a database table. This allows transparency when accessing databases via the document knowledge because all database operations are hidden within the functionality of the document knowledge where they are implemented only once.

The second kind of context information relevant for DAU has already been mentioned: Context in form of possible contents of expected documents. Looking at this from the document-knowledge point-of-view, these expectations are restrictions of more general document types or their parts, respectively. Therefore, incoming expectations are entered as specializations of concepts already defined (e.g., an invoice for a printer of the company HewlettPackard is a specialisation of an invoice of the company HewlettPackard). An example is given in figure 10.

```
expectation Exp_124_Artikelname
expectation–of(Papier–Artikelname<message.Artikelname=("CP1.0–63–05L")>)
expectation–id 124;
description "Top–Konzept für einen Layout/Logik–Dokumentteil in Din A4 Papierdokumenten";
state current;
layout
        :name bottom :type float :range [0.0;29.7]
                :unit cm
                :relevance 0.0;
        :name left :type float :range [0.0;21.0]
                :unit cm
                :relevance 0.0;
        :name page :type integer :range [1;15]
                :relevance 0.0;
        :name right :type float :range [0.0;21.0]
                :unit cm
                :relevance 0.0;
        :name top :type float :range [0.0;29.7]
                :unit cm
                :relevance 0.0;
image
        :name background–color :type string :values (white)
                :relevance 0.0;
        :name foreground–color :type string :values (black)
                :relevance 0.0;
logic
        :name pattern :type string
                :phrase–rules (":pattern "?artikelname (
                        :tolerance 0.65 )"")
                :relevance 0.0

                        :result–locations ("message.Artikelname = ?artikelname");
message
        :name Artikelname :type string :values ("CP1.0–63–05L")
```

Figure 10: Screenshot of the document knowledge browser for an expectation of an article name

Here, we have an expectation for a document part dealing with the name of an article ordered. This frame type is called *expectation*. The only entry in which this expectation differs from the more general concept `Artikelname` (not shown here) is the name of the article in the last row `CP1.0-63-05L` which is the name of a specific cooling element.

Now imagine, an analysis component has generated a result for the analysis of this document part. This result is stored within a frame called *instance* as shown in figure 11. Within the source

```
instance IIExp_124_Artikelname_2_2
        instance–of Exp_124_Artikelname;
        source ("1");
        based–upon (IPapier–Geschäftsbrief_2);
        certainty 1.0;
specialist–name PatternMatcher;
state current;
layout
        :name bottom :unit cm
                :certainty 0.0
                :relevance 0.0  :value 14.596533;
        :name left :unit cm
                :certainty 0.0
                :relevance 0.0  :value 7.5692;
        :name page :certainty 0.0
                :relevance 0.0  :value 1;
        :name right :unit cm
                :certainty 0.0
                :relevance 0.0  :value 9.211734;
        :name top :unit cm
                :certainty 0.0
                :relevance 0.0  :value 14.376399;
image
        :name background–color
                :certainty 0.0
                :relevance 0.0  :value white;
        :name foreground–color
                :certainty 0.0
                :relevance 0.0  :value black;
logic
        :name pattern :certainty 0.8333333
                :relevance 0.0
                        :value "/home/demo/Specialists/ProductDataExtraction/output/1.product";
message
        :name Artikelname
                :certainty 0.8333333
                :relevance 1.0  :value "CP1.0–63–05L";
```

Figure 11: Screenshot of the document knowledge browser for an instance of an article name

slot, a unique number for the document image analysed is given (here number 1), the name of the analysis component is pattern-matcher, and within the message slot `Artikelname` (last few rows), we see that the cooling element CP... has been matched with a certainty of 0.833.

The concrete combination of information coming from expectations and information coming from the corresponding, more general concept depends on the individual DAU component. For some components, e.g, the logo recognition, an expectation is just a restriction of the result space, for others (e.g., the pattern matching component), the expectation is to be unified with the concept (see also the example in chapter 8).

Up to now, we have presented a rough overview on the functionalities of our document knowledge. The following chapters will deal with how this representation is used during analysis.

6 Analysis control

The DAU system is triggered in two different situations: In the first situation, the post-box server triggers DAU control with a new incoming document. In such a case, a process assignment has to be accomplished. The DAU control just retrieves the corresponding plan which denotes a sequence of DAU specialists along with pre- and postconditions and alternative paths. Figure 12 shows a simple analysis plan for the extraction of typical invoice data[1]. With the aid of a cor-

```
plan-1:Plan( name="ExtractAdditionalInvoiceData", next={ start-1 })

start-1:Start( next={ specialist-1 })
ende-1:Ende( prev={ decider-1 })
ende-2:Ende( prev={ specialist-5 })

specialist-1:Specialist( name="ProductDataExtraction", next={ specialist-2 }, prev={ start-1 },
    infoitem={ Papier-Geschäftsbrief.Zahlungsangaben.Preisangaben })
specialist-2:Specialist( name="SenderExtraction", next={ specialist-3 }, prev={ specialist-1 })
specialist-3:Specialist( name="SenderEvaluation", next={ entscheider-1 }, prev={ specialist-2 })
specialist-4:Specialist( name="ShortSenderExtraction", next={ specialist-6 }, prev={ entscheider-1 })
specialist-5:Specialist( name="ResultGeneration", next={ ende-1 }, prev={ entscheider-1 })
specialist-6:Specialist( name="ResultGeneration", next={ ende-2 }, prev={ specialist-4 })

decider-1:Decider( if="Sender rating satisfying?", then={ specialist-5 }, else={ specialist-4 },
                   next={ specialist-5, specialist-4 })
```

Figure 12: A plan for the extraction of invoice data

responding resource file, each specialist can be constructed on the basis of a generic DAU component. Therefore, a resource denotes which specialist may extract which kind of information in a declarative way by using very general paths formulated in the document knowledge. In addition, the resource contains necessary parameters, paths within the document knowledge to be investigated, hardware restrictions and so on. An example is given in figure 13 where the left-hand side shows the program path and the program parameters while the right-hand side shows the naming conventions for the document knowledge, the information items to be ana-

1. Since this plan is only to be invoked after a process assignment it only contains specialists for the extraction of textual data. Whenever this plan is invoked, layout and OCR results are already represented in the document knowledge.

```
[ProductDataExtraction]
program = "/project/vo/bin/pmknow"      DPI-VALUE = 300
slot = "-s"                             LAYOUT-CREATOR = "vote"
pfad = "-p"                             WRITE-ONLY-VARIABLES = false
documentId = "-d"                       TRAPEZOID-WEIGHT = 0.8
expectation = "-e"                      SPECIALIST-NAME = "PatternMatcher"
accuracy = "-a"                         EXP-STRATEGY = "normal"
fast = "-f"                             PATTERN-SLOTGROUP = "logic"
section = "-n"                          PATTERN-SLOT = "pattern"
monitorLevel = "-l"                     PATTERN-FACET = "phrase-rules"
monitorChannel = "-c"                   RANGE-FACET = "range"
feedbackMessage = "-m"                  PAGE-SLOT = "layout.page"
feedbackInstances = "-i"                LEFT-SLOT = "layout.left"
feedbackChannel = "-k"                  OCR-SPECIALIST-NAME = "WORDCLASSIFIER1"
resourceFile = "-r"                     CONCEPT-PARTS = ( "Papier-Geschäftsbrief.Zahlungsangaben.Preisangaben"
orbHost = "-x"                          "Papier-Geschäftsbrief.Leistungsangaben.Warenangaben.
orbPort = "-y"                          Warenbezeichnung.Artikelname" )
remote = "serv-401"
...
```

Figure 13: Partial resource entry for the extraction of product data

lyzed (concept-parts) and some component-specific declarative knowledge.

Using resources and components, specialists are invoked and controlled according to plans. Of course, specialists which have already been executed for the document under consideration are filtered out since their results are already available. Having executed the analysis plan, the DAU control transfers the matching expectation id to the workflow management system. Hence, the document is assigned to this expectation and the corresponding context unit. If requested, additional extracted information is retrieved from the document knowledge (by making usage of its inheritance mechanisms) and handed over to the workflow.

In the second situation, DAU control is invoked when the workflow asks for additional information from a document which has already been assigned to the process. Such a new information extraction task is also specified by macros (e.g., the plan name in figure 12 is such a macro name). In this case, the analysis control retrieves a general information extraction plan and instantiates it. That means, that all specialists are invoked with a restricted task which is far more efficient. As an example, see again the example plan where the specialist for product data extraction shall retrieve the prices of products (`Preisangaben`). The corresponding resource entry in figure 13 shows that this specialist would also be able to extract article names, but in the case of the example plan, this extraction is not carried out.

Our analysis control is visualized by a DAU control post which allows the visualization and inspection of document knowledge, expectations, plans, and document images. Furthermore, parameter settings such as precision or time requirements can be set up here. Figure 14 shows the starting window of the control post. For each document, the analysis plan can be inspected or started in a trigger mode. In this case, a separate window shows all necessary information

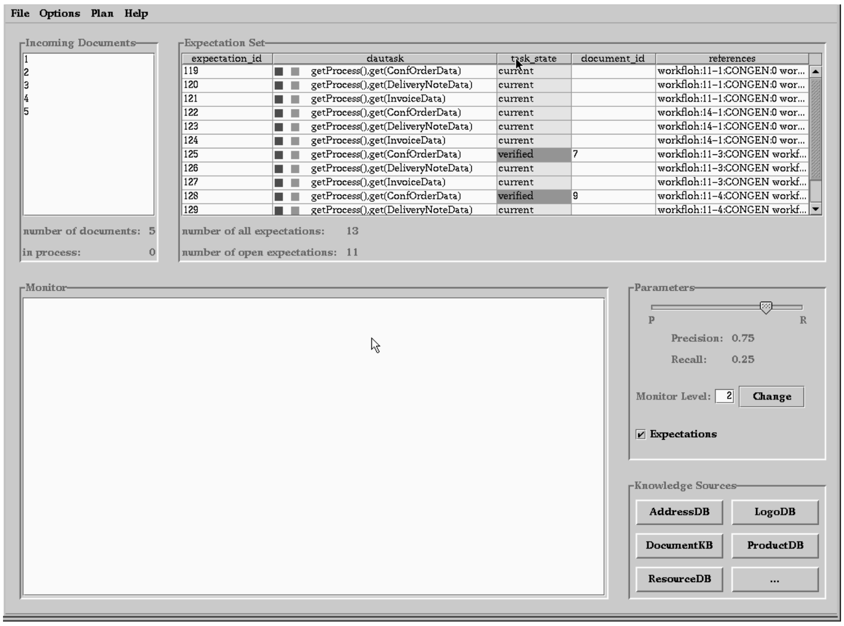


Figure 14: Screenshot of the analysis control post

(see also figure 15).

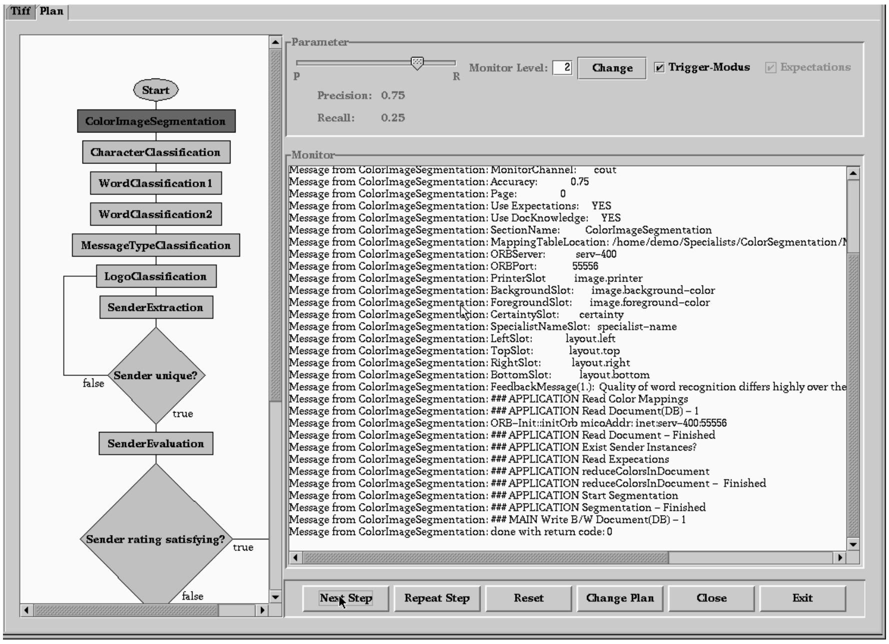


Figure 15: Screenshot of a plan execution

7 DAU components and DAU specialists

All DAU components of the VOPR system are displayed in figure 16. Those components which heavily rely on declarative document knowledge may be transformed into different domain- and task-specific specialists. For such components, the resulting specialists currently employed are shown on the right hand side of figure 16. Now a short description of each component follows: **color image segmentation:** Starting with a color reduction step, the component generates a contrast representation which shows significant color differences in adjacent pixels. It is used to construct a color connected component hierarchy on the basis of the single-pass algorithm. Subsequently, scanner errors are reduced by using typical scannerprofiles. Finally, the color im-

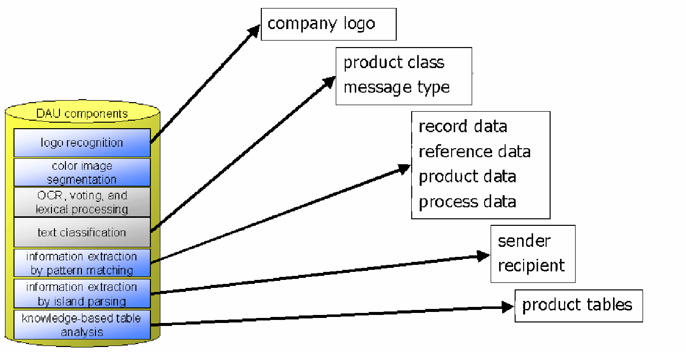


Figure 16: DAU components (left-hand side) and resulting specialists (right-hand side)

age is converted into a black/white image.

logo recognition: The logo recognizer is a by-product of the color image segmentation. First, graphic elements in appropriate document regions are selected as logo candidates. With that, the list of valid logos is filtered by comparing simple features (moments, numbers of colors). The shape of the remaining candidates is compared by recursive tree-matching. Output is a valued ranking of possible logos. Our logo database contains at the moment more than 200 different logos.

text classification: For the textual classification of a document, we employ an inductive rule learner. It learns patterns and boolean expressions on word level during the learning phase and uses fuzzy matching for these expressions in the classification phase. Its output is a valued ranking of matching classes.

OCR, voting and lexical processing: The output of three commercial OCR engines (M/Text, TextBridge and EasyReader) is stored in a complex graph structure. Based on that, our voting component combines character hypotheses by comparing word and line segments and by matching corresponding character graphs. Afterwards, the lexical processing component matches lists, prefixes, suffixes and regular expressions for valid words against the voting results and calculates confidence measures for the final word hypotheses.

pattern matcher: This component allows an error-tolerant, but shallow information extraction. Regular expressions for syntactic text patterns are processed in two steps whereby confidence measures from lexical processing (bottom-up) and from document knowledge (top-down) are combined. The component uses different similarity measures for words based on morphology (word stems, part-of-speech), semantics (synonyms, hypernyms), geometry, and fonts. The pattern matcher generates results for layout and message slotgroups of the document knowledge.

parser: This component accomplishes a deep syntactic analysis for those documents parts which have a strong internal structure. Its kernel is an island parser with a stochastic grammar. The parser generates results for logic and message slotgroups of the document knowledge.

knowledge based table analysis: This component analyses tables in documents which are instantiations of previously defined table models. Analysis is based on the proper detection of a table header by different features (geometry, lines, keywords,...). As a result, the table's structure as well as its contents on cell level are extracted.

For more information on these components, see [24], [25], [26], [27]. There is another component which is not a true analysis component but typically included at the end of a a plan:

result generator: The result generator combines instances produced by other specialists and stored in the document knowledge to a final instance for the whole document image. It is mainly a search algorithm with uncertainty propagation (combination is based on a procedure similar to the MYCIN expert system). Search is based on a chart-parsing approach which allows a high flexibility in the search strategy.

8 Analysis strategies

Those components which can be transformed into several specialists are applicable in situations either with or without expectations. The current situation is submitted to them when being invoked. When dealing with expectations, several strategies can be used: The *closed world strategy* restricts the specialists‘ applications only to expectations. That means that only results which instantiate expectations are produced. The second, *well-rounded strategy* allows results which are consistent with expectations while the last, *comprehensive strategy* allows the generation of both, results based on expectations and results based on more general concepts at the same time.

Figure 17 explains these strategies with respect to the corresponding result space. In addition,

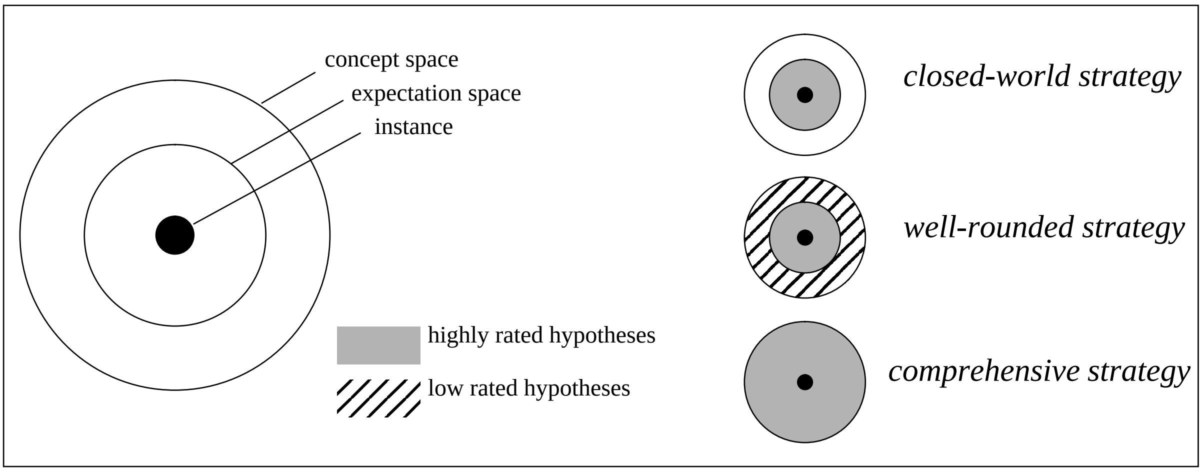


Figure 17: Different analysis strategies and their general effect on results

the well-rounded strategy allows the different weighting of result hypotheses depending on whether they are based on expectations or not.

The strategy used influences the number and kind of analysis instances which are input for the result generator. There is no best strategy because this depends on basic assumptions of the surrounding system environment (How many and what unexpected document may occur beneath those modeled in standard context units?).

We take our pattern matcher as an example for explaining the different strategies: This pattern matcher extracts information items based on text patterns which reveal textual, geometric, and semantic features of a textual expression. Imagine a simple pattern which determines how to extract a price (“total amount:“ `?<NUMBER> currency`). This pattern extracts every number which is enclosed between the string "total amount:" and some currency (currency counts as a hypernym). Imagine also, that there is a price expectation which gives a concrete price (e.g., 105 USD). If running with expectations, the pattern matcher specializes the original pattern to “total amount:“ ?105 “USD“. When using the *closed-world strategy*, the pattern matcher only matches prices which amount to 105 USD. If no price expectation is given by the workflow, the pattern matcher matches no prices at all. However, the threshold for lexical verification of exactly this price might be set comparatively low to avoid rejects. Using the *well-rounded strategy*, the pattern matcher also matches the specialized pattern, however, if no price expectation is given, it matches the general price pattern with a lower a-priori certainty. Using the *comprehensive strategy*, both the specialized and the general pattern are matched.

9 Experimental results

The generation of a proper testbed was a necessary prerequisite in order to prove the usefulness of expectations. It has been quite time-consuming although we could not simulate a realistic scenario with some hundreds of open expectations at a time. Efforts had to be undertaken to simulate one workflow instance for each expectation which must fit to an available paper document, ground-truth data had to be provided and the correct process had to be determined.

We tested the system as a whole up to now with 18 expectations at a time. For process identification, we identified sixteen information items which may be at hand in expectations, e.g, sender, message type, process number, total amount,... However, when simulating the original processes (from which we had all incoming and outgoing documents), we found out that in a typical process, about ten of these items are explicitly mentioned. At the document side, we tested the system with twelve documents which typically contained about six of the relevant information items. The results are shown in table 1.

Table 1: Results for process identification and a few information items with and without expectations

	logo recognition (precision)	logo recognition (recall)	message type (precision)	message type (recall)	references (precision)	references (recall)	Accuracy of process identification
With Expectations	100%	66,6%	90%	75%	94%	80%	100%
Without Expectations	100%	58,3%	80%	66,6%	73,3%	55%	66,6%

This table compares an expectation-guided analysis to a conventional run. Results are shown for some exemplary information items analyzed by different specialists and for the process identification. Not surprisingly, process identification with the usage of expectations was always correct. However, when neglecting expectations for analysis, the final analysis hypothesis led in four cases to a wrong process determination. As we see, the usage of expectations also leads to improvements in precision and in recall. Precision is improved since the number of possible solutions gets smaller, recall is improved since search can accept weaker solutions as correct ones. Moreover, the usage of expectations shortened the runtime of single specialists (e.g., logo recognition, pattern matcher) to a high amount because of the resulting restrictions of the search space.

We also investigated the impact of expectations to DAU components (e.g., in the following for the pattern matching component) in detail. Therefore, we looked at the reasons for errors within 250 documents (when doing a „conventional“ concept-based analysis). Information items analysed were numbers, dates, proper names, and prices. Error rates for these categories ranged from 17%-43%, but the usage of expectations (and especially knowing, e.g., the correct number in advance) can decrease these rates between 20 and 45%. The reason for that is that a lot of errors depended on character recognition problems for single characters (which can be nearly totally repaired by an error-tolerant matching based on expectations).

10 Conclusion

This paper presented a prototypical implementation of a knowledge-based DAU system within an automated business process environment. The VOPR system uses a generic document knowledge representation allowing to incorporate different knowledge sources such as corporate databases and context information from WfMS. Thus, it enables context-driven document analysis and understanding resulting in faster system runs (e.g., reduced search space, extracting only requested data), higher precision (e.g., by using background knowledge), and learning capabilities by incorporating verification results (However, this was not subject of this paper). Additionally, the system incorporates some properties which allow for easy configuration:

- WfMS-vendor independent integration
- free usage of domain ontologies within workflows
- information need of workflows is freely definable
- corporate dispatch strategies via context set
- declarative and flexible formulation of tasks, results, and context
- configurable specialists by component/resource combination
- DAU control based on plans
- model-based document knowledge:
  - domain independency due to models
  - transparent integration of corporate databases
  - inheritance and aggregation provide powerful, though modular concepts

The system accomplishes a tight integration of DAU into WfMS and therefore bridges the gap between paper-based parts of communication and automated business processes. The presented solution for context collection which considers current trends in e-commerce (XML) is efficient while being unintrusive. The only necessary efforts remain at buildtime by including some extra activities in the workflow definition for implementing the integration. The VOPR system is also suitable for requirements of business process reengineering due to its adaptability.

Our future work will further evaluate the impact of expectations on DAU results. On the one hand, we will assess the expenses of efforts for context collection against the revenue in analysis results. And on the other hand, we will estimate the amount of minimal context information necessary to achieve an appropriate assignment rate in real-world quantities of documents.

Acknowledgments

Thanks to our (partly former) colleagues Andreas Lauer, Jürgen Lichter, Michael Malburg, Harald Mayer auf‘m Hofe and Tino Sarodnik for designing and implementing the VOPR system with us. The project has been funded by the German Federal Ministry for Education and Research under grant number 01 IW 807.

References

[1] http://www.kofax.com
[2] http://www.cardiff.com
[3] http://www.captivacorp.com
[4] http://www.serware.de
[5] S. Jablonsky, Ch. Bussler. Workflow Management. *Modeling Concepts, Architecture and Implementation.* International Thomson Computer Press, 1996.
[6] R. Bleisinger, M. Müller, P. Hartmann, T. Dörstling. *Intelligente Eingangspostverarbeitung mit wissensbasierter Dokumentanalyse,* Wirtschaftsinformatik Vol. 41, 1999, pp. 371-377, in German.
[7] T. Bayer. *Understanding Structured Text Documents by a Model-Based Document Analysis System.* Second Int. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan, Oct. 1993, pp. 448-453.
[8] S. Baumann, M. Ben Hadj Ali, A. Dengel, T. Jäger, M. Malburg, A. Weigel, C. Wenzel. *Message Extraction from Printed Documents – A Complete Solution*. Fourth Int. Conf. on Document Analysis and Recognition (ICDAR 97), Ulm, Germany, August 1997.
[9] http://www.im-insiders.de

[10] K.-H. Bläsius, B. Grawemeyer, I. John, N. Kuhn. *Knowledge-Based Document Analysis.* Fourth Int. Conf. on Document Analysis and Recognition, Ulm, Germany, August 1997, pp. 728-731.

[11] http://www.asimus.de

[12] S. W. Lam. *An Adaptive Approach to Document Classification and Understanding*. First workshop on document analysis systems, Kaiserslautern, Germany, 1994, pp. 231-251.

[13] F. Esposito, D. Malerba, G. Semeraro, C. D. Antifora, G. de Gennaro. *Information Capture and Semantic Indexing of Digital Libraries through Machine Learning Techniques*, Fourth Int. Conf. on Document Analysis and Recognition, Ulm, Germany, August 1997, pp. 722-727.

[14] R. Ingold. *A Document Description Language to Drive Document Analysis*, First Int. Conf. on Document Analysis and Recognition, Saint-Malo, France, Sept./ Oct. 2, 1991, pp. 294-301.

[15] U. Remus, F. Lehner. *The Role of Process-oriented Enterprise Modeling in Designing Process-oriented Knowledge Management Systems*. 2000 AAAI Spring Symposium, Bringing Knowledge to Business Processes, Stanford, California.

[16] H. Maus. *Towards a functional integration of document analysis and understanding in workflow management systems*. Proc. of the Workflow Management Conference '99: Workflow-Based Applications, Münster, Germany, Nov. 1999.

[17] E. R. Harold. *XML Bible*. IDG Books Worldwide, Inc, 1999.

[18] http://www.w3c.org/rdf

[19] A. Abecker, A. Bernardi, H. Maus, M. Sintek, C. Wenzel. *Information Supply for Business Processes: Coupling Workflow with Document Analysis and Information Extraction*. Elsevier's Journal of Knowledge-based Systems, Special Issue on Artificial Intelligence in Knowledge Management, Vol. 13, 2000.

[20] http://www.staffware.com

[21] http://herzberg.ca.sandia.gov/jess/

[22] G. Nagy. *What does a machine need to know to read a document?* Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, 1992, pp 1-10.

[23] S. Srihari. *From pixels to paragraphs: the use of models in text recognition.* Second Int. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan, Oct. 1993.

[24] T. Jäger, A. Pies, A.Weigel. *Global versus Local Combination of OCR Results.* Proc. of the 5th JCIS, Feb./ Mar. 2000, Atlantic City, New Jersey, USA.

[25] C. Wenzel, M. Malburg. *Lexicon-Driven Information Extraction from a Document Analysis View*. Int. Workshop on Lexicon Driven Information Extraction, Frascati, Italy, July 1997.

[26] M. Junker, R. Hoch. *An Experimental Evaluation of OCR Text Representations for Learning Document Classifiers*, International Journal on Document Analysis and Recognition, Vol. 1, Number 2, 1998.

[27] C. Wenzel, W. Tersteegen. *Precise Table Recognition by Making Use of Reference Tables.* Seong-Whan Lee, Yasuaki Nakano (Eds.), Document Analysis Systems: Theory and Practice, Lecture Notes in Computer Science, Springer Verlag, Vol. 1655, Aug. 1999.