Pylonix Data Model

Technical Report ifi-2008.06

Department of Informatics
University of Zurich

Christian Tilgner and Dietrich Christopeit
{tilgner,christo}@ifi.uzh.ch

[Version: May 15, 2008]

Abstract

In this report, the data model of the Pylonix approach which is going to be introduced for the first time at ADBIS' 2008 is presented. The purpose of this report is to describe the Pylonix data model and to present the grammar this data model. A complete description of the data model and the grammar is beyond the scope of this report. Detailed information will be presented separately.

Contents

1 Introduction

In this report, the data model of the Pylonix approach which is going to be introduced for the first time in [1] is presented. The purpose of this report is to describe the Pylonix data model and to present the grammar of the data model. A complete description of the data model and the grammar is beyond the scope of this report. Detailed information will be presented separately.

Pylonix is a novel approach for database-based management of complex documents. Its goal is to satisfy the document management request expressed in the Lowell Report [2], by providing all typically available database services also for complex documents and to support functionalities to manage the entire document lifecycle. A document life cycle comprises its creation, storage, manipulation, retrieval and deletion. Pylonix can be integrated in any enterprise architecture that needs fine-grained document management facilities. It offers a data model, which is presented in this report, capable of representing entire documents and it provides database and database model independence. In addition with a novel *Text Query Language (TXQL)* it enables fine-grained manipulation and searching facilities for all document information. TXQL is also beyond the scope of this report and will be presented individually.

After these introductory remarks, the some data model explanations are given in Section 2. Subsequently, the data model grammar as well as attribute definitions are presented in Section 3 and 4.

2 Data Model Description

The Pylonix data model enables the persistent storage of entire documents including all information that belongs to it, such as all complex and multimedia content, their logical and physical structure, as well as various metadata. The metadata includes, for instance, security, data lineage and workflow information, and can be stored on every document layer up to each character.

Together with TXQL it provides complex search functionalities for all document elements. Thus, we offer support for entire documents with a fine granularity and a complex search.

The approach presented in this paper extends the idea of TeNDaX, a collaborative environment for document processing. TeNDaX was for the first time described by Hodel-Widmer and Dittrich in 2004 [3]. The TeNDaX system makes use of a conventional database system managing the entire document life cycle. In TeNDaX, editing text is represented by real-time transactions where every editing action on each character results in one or more database transactions. The system offers security, collaboration, text mining and knowledge management facilities on a very fine-grained level - the individual character.

Our approach seizes this idea from a different point of view and extends it in several ways. Firstly, Pylonix offers fine-grained and complex document management, but is not limited to a collaborative editor as TeNDaX. Secondly, a more powerful data model is used which is capable of representing entire documents with all of its information. TeNDaX as the underlying concept uses an *ordered character sequence (OCS)* as data model for storing documents which suffers from several limitations. It is not capable of modeling all structural information or complex content. On this account, a different data model is necessary to support the full range of document management functionalities of modern word processors.

The novel data model plays a major role in Pylonix. Therefore, we seize the TeNDaX approach of storing each character using an OCS, but extend it to represent complex documents - as known from MS Word or Open Office documents.

Because of the hierarchical nature of a document a tree structure seems to be a feasible representation. Using a tree structure to model a document offers the possibility to realize a fine-grained locking mechanism which allows concurrent access at different levels, i.e. locking nodes at different levels in the tree.

Our Pylonix data model also uses an OCS to store the characters of a document. But instead of using one or two complex trees containing all document information, our model applies multiple small tree structures which are placed upon the OCS. All elements of complex documents have been categorized and assigned to separate trees. In order to avoid unnecessary complexity, the levels of the trees should be kept as small as possible. For our data model we were able to present all information with trees of maximum height four. Separate trees exist for the logical and the physical structure, for layout, lists, tables, figures, links, tables of content, references and fields etc. Except for references and fields, all trees are designed to be independent from each other. With this approach the costs for occurring tree computations are expected to be reduced to a minimum. Moreover, by using such a multiple tree data model, extensibility of the data model is achieved without changing the existing structure. In case of adding further information to the document, new data can be represented by a new tree which can be linked to the OCS.

In Figure 1 an abstract data model for an example document which includes text, structure, several tables, links and figures is depicted. The Pylonix data model is similar to a bridge where the OCS represents the pavement and every type of document information is linked to it like a pylon. With this data model we are able to represent complex documents with all of its complex elements.

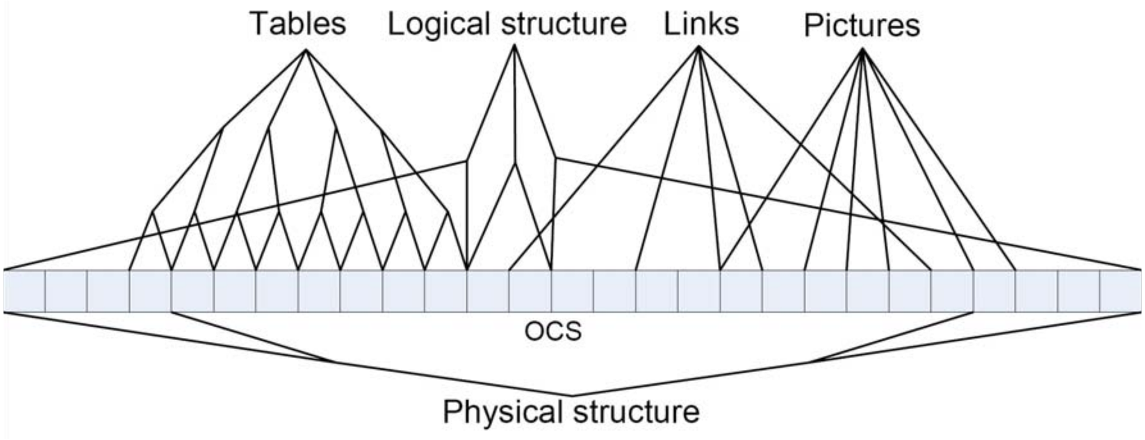

Figure 1: Pylonix Data Model

The belonging Text Query Language is similar to the *Object Query Language (OQL)*. It has to fulfill several requirements. Since the requirement for the support of all document information is defined and finally achieved by the Pylonix data model, the language has to be able to access and to manage all information stored in that data model. Furthermore, TXQL has to be capable of combining all of this document information in their queries in an arbitrary manner. Thus, TXQL enables fine-grained access and a novel complexity of retrieval facilities. All information of complex documents can be accessed, retrieved, manipulated and combined in an arbitrary manner regardless of their e.g. content, structure or metadata.

3 Data Model Grammar

As already mentioned, all elements of complex documents are categorized and assigned to separate tree structures. The grammar of the several tree structures is presented in this section. Additionally, further elements are designed for organizational purposes which are explained herein. A compact version of the grammar is given in Appendix A – Data Model Grammar.

3.1 Tree Structures

The **physical structure** is stored in a tree having a Doc node as its root. Every Doc node has at least one Page node as its children. Pages consist of a header, a main part and footer. The MainPart node again has one or multiple Column nodes as children.

```
PhyDoc          ::= Page {Page}
Page            ::= Header MainPart Footer
MainPart        ::= Column {Column}
```

To model the **logical structure** of a complex document, a tree having a LogDoc node as its root is applied. Each LogDoc node can contain zero or multiple paragraphs followed by zero or multiple chapters. A Chapter node can have one Title followed by zero or multiple Paragraph nodes. Titles again consist of an optional title number (TitleNo) and exactly one title name (TitleName). A paragraph ends with a new line sign.

```
LogDoc          ::= {Paragraph} {Chapter}
Chapter         ::= [Title] {Paragraph}
Title           ::= [TitleNo] TitleName
```

Sections in a complex document are represented by a tree with root node SectionColl and leave nodes called Section. Each SectionColl node has at least one Section leave.

```
SectionColl     ::=  Section {Section}
```

The **footnotes** in a complex document are collected in a tree structure with root node FNColl. Below the root zero or more Footnote nodes are attached. Each Footnote node has two children attached, FNNo and FNText storing information about the footnote number and the actual text content, respectively.

```
FNColl          ::= {Footnote}
Footnote        ::= FNNo  FNText
```

Text is modelled in the style of natural sentences. Thus, below the root node TextColl, Sentence nodes are attached. Each Sentence node has one ore more Word nodes as leaves attached. Words are separated by delimiters such as spaces etc.

```
TextColl        ::= {Sentence}
Sentence        ::= Word {Word}
```

To model the **zone** concept a tree representation is chosen, that collects all available zones in a complex document. These are leaves under the ZoneColl root node describing

- a locked editable region: LockZone
- access rights: SecurityZone
- general notes: NotesZone
- workflow instructions: WorkflowZone
- semantic settings and descriptions (e.g. to support RDF/S): SemanticZone
- visual representation through layout information: LayoutZone
- trust information for reflecting the editors' confidence in a certain text snippet: TrustZone
- the validity of a certain zone: ValidityZone

```
ZoneColl        ::= {Lock | Security | Notes | Workflow | Semantic |
                    Layout | Trust | Validity}
```

Lists are collected under the root node ListColl. Each of the attached List nodes consist of one or more ListItem leave nodes

```
ListColl        ::= {List}
List            ::= ListItem {ListItem}
```

Tables of a complex document are stored in a tree with root node TableColl. Each Table node under the root has one or more Row nodes associated. Accordingly each Row node has one or more associated Column nodes. Nested tables are modelled as separate tables. From the position of the nested table in the OCS it is possible to identify the nested table and its position in the surrounding table.

```
TableColl       ::= {Table}
Table           ::= Row {Row}
Row             ::= Column {Column}
```

Figures are stored under the root node FigureCollection in zero or more the leave nodes Fig.

```
FigureColl      ::= {Fig}
```

Multimedia content like audio and video are stored under the root node AudioColl and VideoColl, respectively. Both root nodes have zero or more AudioElement and VideoElement leave nodes attached, respectively.

```
AudioColl       ::= {AudioElement}
VideoColl       ::= {VideoElement}
```

External linked content referring to an external source is modelled in a LinkCollection root node with attached Link leave nodes.

```
LinkColl        ::= {Link}
```

Descriptions are stored under the root node DescriptionColl that refers to zero or more Description nodes. A Decription node has children DLabel and DText, describing the Label (i.e. a label text like "Figure" and a label number DLNo) and, of course, the caption text itself, respectively.

```
DescriptionColl       ::= {Description}
Caption               ::= DLabel  DText
Label                 ::= DLText  [DLNo]
```

Fields holding automatically generated content is stored under the root node FieldColl. Zero or more Field leave nodes are attached to the root.

```
FieldColl             ::= {Field}
```

Internal linked content referring to internal sources (e.g. a reference to a heading) is modelled in a CrossReferenceColl root node with zero or more attached CrossReference leave nodes.

```
CrossReferenceColl   ::= {CrossReference}
```

The **directories** structure (e.g., table of contents, table of figures etc.) is stored under the DirColl root node that refers to zero or more Dir nodes. These nodes have a DirTitle and zero or more DirItem nodes attached. The DirTitle may have a number, the DirTitleNo attached and must have a heading text, the DirTitle Text. Every DirItem node has a number, the DirItemNo, a text, the DirItemText and a reference to the page the DirItemPage to which the item refers. Depending on the directory DirItemNo and DirItemText refer to different captions, headings or any other content in a document.

```
DirColl               ::= {Dir}
Dir                   ::= DirTitle {DirItem}
DirTitle              ::= [DirTitleNo]  DirTitleText
DirItem               ::= [DirItemNo] DirItemText DirItemPage |
                          DirItemText  [DirItemNo] DirItemPage
```

3.2 Organizational Elements

A **collection of complex documents** is organized under the root node DocCollection. This enables better organization of work that spread across multiple documents, for example of one topic.

```
DocCollection         ::= {Doc}
```

To assign **format templates** for pages, a page format template is linked to a Page node. All page format templates are registered in a page format template collection (PageFormatTemplateColl).

```
PageFormateTemplateColl    ::=     PageFormatTemplate
                                   {PageFormatTemplate}
```

Individual formatting of entire tree structures (e.g., tables etc.) is achieved using layout zones. Within the layout zone a **style sheet** can be referenced defining the general formatting of the tree. The available style sheets for a document are collected under the StylesheetColl root node.

```
StylesheetColl        ::= Stylesheet {Stylesheet}
```

4 Attribute Definitions

In this section a definition and short explanation of the attributes of the data model elements is presented. Besides the tree structures and the organizational elements, the data model elements comprise the chars of the ordered character sequence as well as border elements by which the connection between the leaves of the trees and the OCS is realized. A compact version of the attribute definitions is given in Appendix B – Attribute Definition.

4.1 Tree Structures

Physical Structure

Node	**Attribute**	**Description**
PhyDoc	PageIDs	References to the Page nodes of the document (ordered)
Page	PhyDocID	Reference to its parent node
	HeaderID	Reference to the header of the page
	MainPartID	Reference to the main part of the page
	FooterID	Reference to the footer of the page
	PageFormatTemplateID	Reference to the page format set in the section node
	StartCharID	Reference to the first char of the page
	EndCharID	Reference to the last char of the page
	PageNoPhy	Physical page number
	PageNoLog	Logical page number appearing on the page
	IsPageNoDependend	True if page number is dependent of the page number appeared on the previous page
	IsFirstPageOfSection IsLastPageOfSection Hidden	Important to realize manual section changes
Header	PageID	Reference to its parent node
	StartCharID	Reference to the first char of the header
	EndCharID	Reference to the last char of the header
Footer	PageID	Reference to its parent node
	StartCharID	Reference to the first char of the footer
	EndCharID	Reference to the last char of the footer
MainPart	PageID	Reference to its parent node
	ColumnIDs	References to the page columns (ordered)
	StartCharID	Reference to the first char of the main part
	EndCharID	Reference to the last char of the main part
Column	MainPartID	Reference to its parent node
	StartCharID	Reference to the first char of the column
	EndcharID	Reference to the last char of the column
	ColumnNo	Number of the page column

Logical Structure

Node	Attribute	Description
LogDoc	ChapterIDs	References to the Chapter nodes of the document (ordered)
	ParagraphIDs	References to the Paragraph nodes of the document (ordered)
Chapter	LogDocID	Reference to its parent node
	TitleID	Reference to the title of the chapter
	ParagraphIDs	References to the paragraphs in this chapter
	ParentChapterID	Reference to the chapter it belongs to
	ChildChapterIDs	Reference to subchapters if existing
	StartCharID	Reference to the first char of the chapter
	EndCharID	Reference to the last char of the chapter
	ChapterNo	Chapter number
	Creator	Name of the creator
	TS_Created	Date of creation
Paragraph	ParentID	Reference to the parent node (LogDoc or Chapter)
	StartCharID	Reference to the first char of the paragraph
	EndCharID	Reference to the last char of the paragraph
	Creator	Name of the creator
	TS_Created	Date of creation
Title	ChapterID	Reference to its parent node
	TitleNoID	Reference to the title number of the title
	TitleTextID	Reference to the text of the title
	StartCharID	Reference to the first char of the title
	EndCharID	Reference to the last char of the title
TitleNo	TitleID	Reference to its parent node
	StartCharID	Reference to the first char of the title number
	EndCharID	Reference to the last char of the title number
TitleText	TitleID	Reference to its parent node
	StartCharID	Reference to the first char of the title text
	EndCharID	Reference to the last char of the title text

Sections

Node	Attribute	Description
SecColl	SectionIDs	Reference to the Section nodes (ordered)
Section	SecCollID	Reference to the parent node
	PageIDs	References to the pages that belong to the section
	PageFormatTemplateID_Std	Reference to the node with the page format template applied to the section
	PageFormatTemplateID_Uneven	Same as above – for uneven pages if two sided format is set for the document
	StartCharID	Reference to the first char of the section
	EndCharID	Reference to the last char of the section
	SectionNo	Section number
	Creator	Name of the creator

	TS_Created	Date of creation

Footnotes

Node	Attribute	Description
FNColl	FootnoteIDs	References to the existing footnotes (ordered)
Footnote	FNCollID	Reference to the parent node
	FNNoID	Reference to the child node footnote number
	FNTextID	Reference to the child node footnote text
	Creator	Name of the creator
	TS_Created	Date of creation
FNNo	FootnoteID	Reference to the parent node
	StartCharID	Reference to the first char of the footnote number
	EndCharID	Reference to the last char of the footnote number
FNText	FootnoteID	Reference to the parent node
	StartCharID	Reference to the first char of the footnote text
	EndCharID	Reference to the last char of the footnote text

Text

Node	Attribute	Description
TextColl	SentenceIDs	References to the existing sentences (ordered)
Sentence	TextCollID	Reference to the parent node
	WordIDs	References to the existing words (ordered)
	StartCharID	Reference to the first char of the sentence
	EndCharID	Reference to the last char of the sentence
Word	SentenceID	Reference to the parent node
	StartCharID	Reference to the first char of the word
	EndCharID	Reference to the last char of the word

Zones

Node	Attribute	Description
ZoneColl	LockZoneIDs	References to the existing lock zones
	SecZoneIDs	References to the existing security zones
	NoteZoneIDs	References to the existing note zones
	WFZoneIDs	References to the existing workflow zones
	SemZoneIDs	References to the existing semantic zones
	LayoutZoneIDs	References to the existing layout zones
	TrustZoneIDs	References to the existing trust zones
	ValidityZoneIDs	References to the existing validity zones
LockZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification

	Last_Action	Description of last modification
	Session	True if lock is only for current user session
SecZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	AccessMatrix	Matrix representation of users' access rights, e.g. read, not write etc.
NoteZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	Text	Note
WFZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	Classification	Desired end-status
	Status	Current status this zone is in
	Instruction	Workflow instructions
SemZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	ToDo	
LayoutZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	StylesheetID	Reference to the use style sheet
	Font	Font
	Fontsize	Font size
	Bold	Bold
	Italian	Italian

	Underlined	Underlined
	Doublelined	Double lined
	Canceled	Canceled
	DoubleCanceled	Double canceled
	Inferior	Inferior
	Superior	Superior
	Capitals	Capitals
	FontColor	Font color
	BackgroundColor	Background color
	Alignment	Alignment
	Shift	shifting value
TrustZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	Signature	Digital signature of the zone, verifiable through validation services (certificate validation , signature servers)
ValidityZone	ZoneColl	Reference to the parent node
	StartCharID	Reference to the first char of the zone
	EndCharID	Reference to the last char of the zone
	Creator	Name of the creator
	TS_Created	Date of creation
	TS_LastChanged	Date of last modification
	Last_Action	Description of last modification
	Valid_From	Valid from
	Valid_Until	Valid until

Lists

Node	**Attribute**	**Description**
ListColl	ListIDs	References to the existing lists
List	ListCollID	Reference to its parent node
	ListItemIDs	References to the list items belonging to it (ordered)
	ParentListItemID	Reference to the list item it belongs to, if any
	ChildList[OwnListItemID][RefListID]	Assignment which list item contains which list
	StartCharID	Reference to the first char of the list
	EndCharID	Reference to the first char of the list
	ListType_numbered	Numbered or unnumbered list
	Symbol	Symbol for unnumbered lists
	StartNo	Starting number for numbered lists
	Creator	Name of the creator
	TS_Created	Date of creation
ListItem	StartCharID	Reference to the first char of the list item

	EndCharID	Reference to the first char of the list item

Tables

Node	Attribute	Description
TableColl	TableIDs	References to the existing tables
Table	TableCollID	Reference to its parent node
	RowIDs	Reference to the row nodes belonging to it (ordered)
	StartCharID	Reference to the first char of the table
	EndCharID	Reference to the first char of the table
	RowAmount	Amount of rows
	ColAmount	Amount of columns
	Alignment	Alignment of the table
	Creator	Name of the creator
	TS_Created	Date of creation
Row	TableID	Reference to its parent node
	ColumnIDs	References to the columns of the row (ordered)
	StartCharID	Reference to the first char of the row
	EndCharID	Reference to the first char of the row
	RowNo	Number of that row
	Creator	Name of the creator
	TS_Created	Date of creation
Column	RowID	Reference to its parent node
	StartCharID	Reference to the first char of the column
	EndCharID	Reference to the first char of the column
	ColumnNo	Number of that column
	Creator	Name of the creator
	TS_Created	Date of creation

Figures

Node	Attribute	Description
FigureColl	FigureIDs	References to the existing figures
Figure	FigureCollID	Reference to its parent node
	CharID	Reference to the belonging char in the OCS
	SourceLink	Link to the Source of the figure
	Format	File format
	DescrID	Reference to the belonging description node
	SizeX	Figure width (pixel)
	SizeY	Figure height (pixel)
	PositionX	X-position of the figure on the Page
	PositionY	Y-position of the figure on the Page
	Creator	Name of the creator
	TS_Created	Date of creation

Audio Clips

Node	**Attribute**	**Description**
AudioColl	AudioIDs	References to the existing audio clips
Audio	AudioCollID	Reference to its parent node
	CharID	Reference to the belonging char in the OCS
	SourceLink	Link to the Source of the audio clip
	Format	File format
	Duration	Duration of the audio clip
	DescrID	Reference to the belonging description node
	Creator	Name of the creator
	TS_Created	Date of creation

Video Clips

Node	**Attribute**	**Description**
VideoColl	VideoIDs	References to the existing video clips
Video	VideoCollID	Reference to its parent node
	CharID	Reference to the belonging char in the OCS
	SourceLink	Link to the Source of the video clip
	Format	File format
	Duration	Duration of the video clip
	Solution	Solution of the video clip
	DescrID	Reference to the belonging description node
	SizeX	Video width (pixel)
	SizeY	Video height (pixel)
	PositionX	X-position of the Video on the Page
	PositionY	Y-position of the Video on the Page
	Creator	Name of the creator
	TS_Created	Date of creation

Links

Node	**Attribute**	**Description**
LinkColl	LinkIDs	References to the existing links
Link	LinkCollID	Reference to its parent node
	Destination	Destination of that link
	StartCharID	Reference to the first char of the link
	EndCharID	Reference to the first char of the link
	Creator	Name of the creator
	TS_Created	Date of creation

Description

Node	**Attribute**	**Description**
DescriptionColl	DescriptionIDs	References to the existing descriptions
Description	DescriptionCollID	Reference to its parent node

	DLabelID	Reference to its label node
	DTextID	Reference to the description text node
	StartCharID	Reference to the first char of the description
	EndCharID	Reference to the first char of the description
	DescrType	Type of description, e.g. figure, table, formula
	DescrNo	Description number
	Creator	Name of the creator
	TS_Created	Date of creation
DText	DescrID	Reference to its parent node
	StartCharID	Reference to the first char of description text
	EndCharID	Reference to the first char of description text
DLabel	DescrID	Reference to its parent node
	StartCharID	Reference to the first char of description label
	EndCharID	Reference to the first char of description label
	DLTextID	Reference to the label text node
	DLNoID	Reference to the label number node
DLText	DLabelID	Reference to its parent node
	StartCharID	Reference to the first char of the label text
	EndCharID	Reference to the first char of the label text
DLNo	DLabelID	Reference to its parent node
	StartCharID	Reference to the first char of the label number
	EndCharID	Reference to the first char of the label number

Fields

Node	**Attribute**	**Description**
FieldColl	FieldIDs	References to the existing fields
Field	FieldCollID	Reference to its parent
	FieldType	Type of field (e.g. author, date)
	StartCharID	Reference to the first char of the field
	EndCharID	Reference to the first char of the field
	SourceCopyFrom	ID or attribute of existing element which is copied to this location
	Creator	Name of the creator
	TS_Created	Date of creation

Cross References

Node	**Attribute**	**Description**
CrossRefColl	CrossRefIDs	References to the existing cross references
CrossRef	CrossRefCollID	Reference to parent node
	CrossRefType	Type of Cross Reference (e.g. numbered element, text, footnote, figure, table)
	SourceCopyFrom	ID of existing element whose chars are copied to this location
	StartCharID	Reference to the first char of the reference
	EndCharID	Reference to the first char of the reference
	Creator	Name of the creator

	TS_Created	Date of creation

Directories

Node	Attribute	Description
DirColl	DirIDs	References to the existing directories
Dir	DirCollID	Reference to its parent node
	DirTitleID	Reference to the directory title
	DirItemIDs	Reference to the directory items (ordered)
	StartCharID	Reference to the first char of the directory
	EndCharID	Reference to the first char of the directory
	DirType	DirectoryType (e.g. Table of content / figures)
	TotalLayerAmount	Amount of layers
	FillChar	code of fill character
	Creator	Name of the creator
	TS_Created	Date of creation
DirTitle	DirID	Reference to the directory it belongs to
	DirTitleNoID	Reference to the title number node
	DirTitleTextID	Reference to the title text node
	StartCharID	Reference to the first char of the directory title
	EndCharID	Reference to the first char of the directory title
DirTitleNo	DirTitleID	Reference to its parent node
	StartCharID	Reference to the first char of the title number
	EndCharID	Reference to the first char of the title number
DirTitleText	DirTitleID	Reference to its parent node
	StartCharID	Reference to the first char of the title text
	EndCharID	Reference to the first char of the title text
DirItem	DirID	Reference to the directory it belongs to
	DirItemNoID	Reference to the item number node
	DirItemTxtID	Reference to the item text node
	DirItemPgID	Reference to the item page node
	StartCharID	Reference to the first char of the directory item
	EndCharID	Reference to the first char of the directory item
	LayerNo	Layer number of that item
DirItemNo	DirItemID	Reference to its parent node
	StartCharID	Reference to the first char of the item number
	EndCharID	Reference to the first char of the item number
	SourceCopyFrom	ID of existing element whose chars are copied to this location
DirItemText	DirItemID	Reference to its parent node
	StartCharID	Reference to the first char of the item text
	EndCharID	Reference to the first char of the item text
	SourceCopyFrom	ID of existing element whose chars are copied to this location
DirItemPg	DirItemID	Reference to its parent node
	StartCharID	Reference to the first char of the item page
	EndCharID	Reference to the first char of the item page

	SourceIDForComputation	ID of references element – used to compute the actual page number

Formulas

Node	**Attribute**	**Description**
FormulaColl	FormulaIDs	References to the existing formulas
Formula	FormulaCollID	Reference to its parent
	To Do	

4.2 Organizational Elements

Document Collection

Node	**Attribute**	**Description**
DocColl	DocIDs	References to the existing documents
	DocCollName	Name of document collection
	Creator	Creator of document collection
	TS_Created	Date of creation
	Last Access	Last accessed by which user
	Group	User group
	Security	Collection access permissions

Documents

Node	**Attribute**	**Description**
Doc	PhyDocID	References to the tree with its physical structure
	LogDocID	References to the tree with its logical structure
	ZoneCollID	References to the tree with its zone information
	SecCollID	References to the tree with its sections
	TextCollID	References to the tree with its text
	ListCollID	References to the tree with its lists
	FNCollID	References to the tree with its footnotes
	TableCollID	References to the tree with its tables
	FigureCollID	References to the tree with its figures
	LinkCollID	References to the tree with its links
	DescrCollID	References to the tree with its descriptions
	FieldCollID	References to the tree with its fields
	CrossRefCollID	References to the tree with its cross references
	DirCollID	References to the tree with its directories
	FormulaCollID	References to the tree with its formulas
	StartCharID	Reference to the first char of the document
	EndCharID	Reference to the first char of the document
	Creator	Creator of document
	DocName	Document name
	Authors	All authors sorted by time of last authoring
	TS_Creation	Date of creation
	TS_LastChanged	Date of last modification
	LastAction	Type of last modification
	TwoSided	True if two sided document format is applied
	TrackChanges TS_TrackChangesSince TrackChangesInitiator	Relevant for change notification
	Group	User group
	Security	Global document access permissions

Page Format Templates

Node	Attribute	Description
PageFormatTemplateColl	PageFormatTemplateIDs	Reference to existing templates
PageFormatTemplate	PageFormatTemplateCollID	Reference to its parent
	TemplateName	Name of template
	Landscape	True if landscape format is applied
	TwoSided	True if two sided document format is applied
	BorderTop	Height of upper border
	BorderBottom	Height of lower border
	FooterHeight	Height of footer
	SpaceFooterMain	Space between footer and main part
	HeiderHeight	Height of header
	SpaceHeaderMain	Space between header and main part
	ColumnAmount	Amount of page columns
	ColumnWidth	Ordered Array of column widths
	SpaceBetweenColumns	Ordered Array of space between the single columns
	PageSizeX_portrait	Page width for portrait format
	PageSizeX_landscape	Page width for landscape format
	PageSizeY_portrait	Page height for portrait format
	PageSizeY_landscape	Page height for landscape format
	MainPartHeight_portrait	Height of main part of page for portrait format
	MainPartHeight_landscape	Height of main part of page for landscape format
	BorderLeft_one	Width of left page border for one sided page format
	BorderLeft_two	Width of left page border for two sided page format
	BorderRight_one	Width of right page border for one sided page format
	BorderRight_two	Width of right page border for two sided page format

Style sheets

Node	Attribute	Description
StyleSheetColl	StyleSheetIDs	References to the existing style sheets
Stylesheet	StyleSheetCollID	Reference to its parent
	StyleSheetName	Name of style sheet
	Creator	Name of the creator
	Font	Font
	Fontsize	Font size
	Bold	Bold

	Italian	Italian
	Underlined	Underlined
	Doublelined	Double lined
	Canceled	Canceled
	DoubleCanceled	Double canceled
	Inferior	Inferior
	Superior	Superior
	Capitals	Capitals
	FontColor	Font color
	BackgroundColor	Background color
	Alignment	Alignment
	Shift	shifting value

4.3 Ordered Character Sequence

Chars

Node	Attribute	Description
Char	PredecessorID	References to its predecessor char
	SuccessorID	Reference to its successor car
	BEID	Reference to its border element
	Value	Value of the char
	Creator	Name of creator
	TS_Creation	Date of creation
	TS_LastActions	Dates of last modification (ordered)
	Authors_LastUsed	Authors who did last modifications (ordered)
	LastActions	Descriptions of last modifications (ordered)

4.4 Border Elements

Border Elements

Node	Attribute	Description
BorderElement	CharID	References to the char it belongs to
	Refs[Type][StartID][EndID]	Array storing references to the tree nodes

References

[1] Tilgner, C., Christopeit, D., Dittrich, K. R., Ziegler, P.: Pylonix: A Database Module for Collaborative Document Management. To Appear In: ADBIS '08, Pori, Finland. (2008).

[2] Abiteboul, S., et al.: The Lowell Database Research Self Assessment. CoRR. 0310006, 2003.

[3] Hodel, T.B., Dittrich K.R.: Concept and Prototype of A Collaborative Business Process Environment for Document Processing. IEEE TKDE, 52(1), 61-120, (2005).

Appendix A – Data Model Grammar

Tree Structures

Physical Structure

```
PhyDoc          ::= Page {Page}
Page            ::= Header MainPart Footer
MainPart        ::= Column {Column}
```

Logical Structure

```
LogDoc          ::= {Paragraph} {Chapter}
Chapter         ::= [Title] {Paragraph}
Title           ::= [TitleNo] TitleName
```

Sections

```
SectionColl     ::=  Section {Section}
```

Footnotes

```
FNColl          ::= {Footnote}
Footnote        ::= FNNo  FNText
```

Text

```
TextColl        ::= {Sentence}
Sentence        ::= Word {Word}
```

Zones

```
ZoneColl        ::= {Lock | Security | Notes | Workflow | Semantic |
                       Layout | Trust | Validity}
```

Lists

```
ListColl        ::= {List}
List            ::= ListItem {ListItem}
```

Tables

```
TableColl           ::= {Table}
Table               ::= Row {Row}
Row                 ::= Column {Column}
```

Figures

```
FigureColl          ::= {Fig}
```

Multimedia Content

```
AudioColl           ::= {AudioElement}
VideoColl           ::= {VideoElement}
```

External Linked Content

```
LinkColl            ::= {Link}
```

Descriptions

```
DescriptionColl     ::= {Description}
Caption             ::= DLabel  DText
Label               ::= DLText  [DLNo]
```

Fields

```
FieldColl           ::= {Field}
```

Internal Linked Content

```
CrossReferenceColl  ::= {CrossReference}
```

Directories

```
DirColl             ::= {Dir}
Dir                 ::= DirTitle {DirItem}
DirTitle            ::= [DirTitleNo]  DirTitleText
DirItem             ::= [DirItemNo] DirItemText DirItemPage |
                        DirItemText  [DirItemNo] DirItemPage
```

Organizational Elements

Collection of Complex Documents

```
DocCollection               ::= {Doc}
```

Page Format Templates

```
PageFormateTemplateColl     ::= PageFormatTemplate
                                {PageFormatTemplate}
```

Style Sheets

```
StylesheetColl       ::= Stylesheet {Stylesheet}
```

Appendix B – Attribute Definitions

Tree Structures

Physical Structure

PhyDoc:	PageIDs (ordered)
Page:	PhyDocID, HeaderID, MainPartID, FooterID, PageFormatTemplateCloneID, StartCharID, EndcharID, PageNoPhy, PageNoLog, IsPageNoDependend, IsFirstPageOfSection, IsLastPageOfSection, Hidden
Header	PageID, StartCharID, EndCharID
Footer	PageID, StartCharID, EndCharID
MainPart	PageID, ColumnIDs (ordered), StartCharID, EndCharID
Column	MainPartID, StartCharID, EndcharID, ColumnNo

Logical Structure

LogDoc	ChapterIDs (ordered), ParagraphIDs (ordered)
Chapter	LogDocID, TitleID, ParagraphIDs (ordered), ParentChapterID, ChildChapterIDs (ordered), StartCharID, EndCharID, ChapterNo, Creator, TS_Created
Paragraph	ParentID, StartCharID, EndcharID, Creator, TS_Created
Title	ChapterID, TitleNoID, TitleTextID, StartCharID, EndCharID,
TitleNo	TitleID, StartCharID, EndCharID,
TitleText	TitleID, StartCharID, EndCharID,

Sections

SecColl	SectionIDs (ordered)
Section	SecCollID, PageIDs, PageFormatTemplateID_Std, PageFormatTemplateID_Uneven, StartCharID, EndCharID, SectionNo, Creator, TS_Created

Footnotes

FNColl	FootnoteIDs (ordered)
Footnote	FNCollID, FNNoID, FNTextID, Creator, TS_Created
FNNo	FootnoteID, StartCharID, EndCharID
FNText	FootnoteID, StartCharID, EndCharID

Text

TextColl	SentenceIDs(ordered)
Sentence	TextCollID, WordIDs(ordered), StartCharID, EndCharID,
Word	SentenceID, StartCharID, EndCharID

Zones

ZoneColl	LockZoneIDs, SecZoneIDs, NoteZoneIDs, WFZoneIDs, SemZoneIDs, LayoutZoneIDs, TrustZoneIDs, ValidityZoneIDs

Common attributes of all zones:

*Zone	ZoneCollID, StartCharID, EndCharID, Creator, TS_Creation, TS_LastChanged, Last_Action

Specific zone attributes:

LockZone	Session
SecZone	AccessMatrix
NoteZone	Text
WFZone	Classification, Status, Instruction
SemZone	ToDo
LyZone	StylesheetID, Font, Fontsize, Bold, Italian, Underlined, Doublelined, Canceled, Double Canceled, Inferior, Superior, Capitals, FontColor, BackgroundColor, Alignment, Shift
Trust	ToDo
Validity	ToDo

Lists

ListColl	ListIDs
List	ListCollID, ListItemIDs (ordered), ParentListItemID, ChildListIDs [ListItemID,ChildListID], StartCharID, EndCharID, ListType_numbered, Symbol, StartNo, Creator, TS_Created
ListItem	StartCharID, EndcharID

Tables

TableColl	TableIDs
Table	TableCollID, RowIDs (ordered), RowAmount, ColAmount, StartCharID, EndCharID, Alignment, Creator, TS_Created
Row	TableID, ColumnIDs (ordered), StartCharID, EndCharID RowNo, Creator, TS_Created
Column	RowID, StartCharID, EndCharID, ColumnNo, Creator, TS_Created

Figures

FigureColl	FigureIDs
Figure	FigureCollID, CharID, SourceLink, Format, DescrID, SizeX, SizeY, PositionX, PositionY, Creator, TS_Created

Audio Clips

AudioColl	AudioIDs
Audio	AudioCollID, CharID, SourceLink, Format, Duration, DescrID, Creator, TS_Created

Video Clips

VideoColl	VideoIDs
Video	VideoCollID, CharID, SourceLink, Format, Duration, Solution, DescrID, SizeX, SizeY, PositionX, PositionY, Creator, TS_Created

Links

LinkColl	LinkIDs
Link	LinkCollID, Destination, StartCharID, EndCharID, Creator, TS_Created

Description

DescrColl	DescrIDs
Descr	DescrCollID, DLabelID, DTextID, StartCharID, EndCharID, DescrType, DescrNo, Creator, TS_Created
DText	DescrID, StartCharID, EndcharID
DLabel	DescrID, StartCharID, EndCharID, DLText, DLNo

DLText	DLabelID, StartCharID, EndcharID
DLNo	DLabelID, StartCharID, EndcharID

Fields

FieldColl	FieldIDs
Field	FieldCollID, StartCharID, EndcharID, FieldType, SourceCopyFrom, Creator, TS_Created

Cross References

CrossRefColl	CrossRefIDs
CrossRef	CrossRefCollID, StartCharID, EndcharID, CrossRefType, SourceCopyFrom, Creator, TS_Created

Directories

DirColl	DirIDs
Dir	DirCollID, DirTitleID, DirItemIDs (ordered), StartCharID, EndcharID, DirType, TotalLayerAmount, FillChar, Creator, TS_Created
DirTitle	DirID, DirTitleNoID, DirTitleTextID, StartCharID, EndcharID
DirTitleNo	DirTitleID, StartCharID, EndcharID
DirTitleText	DirTitleID, StartCharID, EndcharID
DirItem	DirID, DirItemNoID, DirItemTxtID, DirItemPgID, StartCharID, EndcharID, LayerNo
DirItemNo	DirItemID, StartCharID, EndCharID, SourceCopyFrom
DirItemTxt	DirItemID, StartCharID, EndCharID, SourceCopyFrom
DirItemPg	DirItemID, StartCharID, EndCharID, SourceIDForComputation

Formulas

FormColl	::=	Formula
Formula	::=	ToDo

Organizational Elements

Document Collection

DocColl	DocIDs, DocCollName, Creator, TS_Created, LastAccess, Group, Security

Documents

Doc	PhyDocID, LogDocID, ZoneCollID, SecCollID, TextCollID, ListCollID, FNCollID, TableCollID, FigureCollID, LinkCollID, DescrCollID, FieldCollID, CrossRefCollID, DirCollID, FormulaCollID, StartCharID, EndCharID, Creator, DocName, Authors, TS_Creation, TS_LastChanged, LastAction, OneTwoSided, TrackChanges, TS_TrackChangesSince, TrackChangesInitiator Group, Security

Page Format Templates

PageFormateTemplatColl	DocID, PageFormatTemplateIDs
PageFormatTemplate	PageFormatTemplateCollID, TemplateName, Landscape, TwoSided, BorderTop, BorderBottom, FooterHeight, SpaceFooterMain, HeaderHeight, SpaceHeaderMain ColumnAmount, ColumnWidths (ordered Array) SpaceBetweenColumns (ordered Array) PageSizeX_portrait, PageSizeX_landscape PageSizeY_portrait, PageSizeY_landscape, MainPartHeight_portrait, MainPartHeight_landscape BorderLeft_one, BorderLeft_two BorderRight_one, BorderRight_two

Style sheets

StyleSheetColl	StyleSheetIDs
StyleSheet	StyleSheetCollID, StyleSheetName, Creator, Font, Fontsize, Bold, Italian, Underlined, Doublelined, Canceled, Double Canceled, Inferior, Superior, Capitals, FontColor, BackgroundColor, Alignment, Shift

Ordered Character Sequence

Chars

Char	PredecessorID, SuccessorID, BEID, Value, Creator, TS_Creation, TS_LastActions, Authors_LastActions, LastActions

Border Elements

Border Elements

BorderElement	CharID Refs[Type][StartID][EndID]