An algebra for structured documents in the context of the object-oriented approach
Sylviane Schwer, Haider Hamza, André Flory

To cite this version:
Sylviane Schwer, Haider Hamza, André Flory. An algebra for structured documents in the context of the object-oriented approach. 19 pages. 1999. <hal-00265817>

HAL Id: hal-00265817
https://hal.archives-ouvertes.fr/hal-00265817
Submitted on 20 Mar 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
An algebra for structured documents in the context of the object-oriented approach

Haider Hamza
INSA, L3I 20, Avenue Albert Einstein 69621 Villeurbanne France
e-mail : hamza@lisflury.insa-lyon.fr

Sylviane R. Schwer
Université Paris 13, LIPN, av. J. B., Clément, 93430 Villetaneuse, France
e-mail: schwer@lipn.univ-paris13.fr

André Flori
INSA, L3I 20, Avenue Albert Einstein 69621 Villeurbanne France
e-mail: flori@lisflury.insa-lyon.fr

Summary

The aim of this paper consists in defining an algebra allowing the request of structured documents in the context of object-oriented approach. The operators of this algebra are defined in conformity with the concepts of the model object. It takes into consideration not only the embedding of structures and links between them, but also the informational aspect of the document as well as its structural aspect.

The operands of the proposed algebra are only one type: they are substructures. A substructure is made of a collection of the databases objects, grouped in classes and interconnected through links. This algebra responds to two objectives. Firstly, it represents a basic nucleus of a declarative query system; it contains the whole of the elementary operations which will be used in the resolution of the request. Secondly, it provides facilities to users to formulate their requests and manipulate the documentary database.

Key Words : Structured documents, Object-oriented, Documentary database, Query, Algebra

1. Introduction

Since the emergence of the relational model, many researches have addressed the problem of the development of documentary databases using the technology of database management relational systems (RDBMS). For the development of documentary applications, the interest of this approach lies in the possibility of taking advantage of all the functionality of RDBMS. Nevertheless the capacities of abstraction and the expressiveness of the relational model are insufficient for handling structured documents. Indeed, the representation of the fragments of the documents by several tables results in a heavy modeling and decreases considerably the performance of the system.

Researches aiming at a new generation of DBMS has been carried out to extend the models of data representation and to increase the power of manipulation languages. Indeed, new needs have emerged requiring references to the stocked information in terms of the perception of the user and not simply according to their model inside the database. The object oriented DBMS gives nowadays a quite rich model capable of representing complex data. A document can be modeled by a complex object the components of which may be formed by other objects belonging to different classes. The classes are organized in a hierarchy of inheritance in which the objects of different classes respect the relations of inclusion between them. The technology of OODBMS is well adapted to the management of structured documents such as arborescences of objects with associated methods [Amgh89, Ben91, Mou91, Hau92].

The retrieval of informations is one of the most important functionality of a management system of documentary databases. By system of retrieval of information, we mean the whole mechanisms which allows the user to select documentary information. One of the essential objectives of a retrieval system of informations is to make easy the restitution of a portion of information from a documentary database, in response to a user request.

Different strategies of information retrieval were developed, and most of them were built around an algebra. In fact, G. ting [Gu89] has proposed an algebra made of a set of operators which permits to take into account the different informational aspects of a document. But this algebra is not well adapted to an object-oriented system, because it doesn't consider the inter-objects relationships. In the case of the algebra for the OODBMS, several works were carried out [Alha93, Cue90, Hau93, Lin93, Sun90, Sha90, Sub93]. However, these algebra are not really adapted to structured documents, since they only take into consideration the informational aspect of the data, and not at all their structural aspect. This is the case, for instance, of the algebra of Show and Zdonik, which is an extension of the algebra of the denormalized models that takes into account some object-oriented concepts. The EXCCESS algebra [vand91], is a many-sort algebra i.e. the operators are defined by types of operators. Such an approach in a documentary retrieval context puts in question the uniformity of the operators and the reusing of the results of the requests for a new query. The algebra of Lu [Lu93] is based on a clear distinction between links of association and links of aggregation. We think that in fact this distinction is not relevant: two concepts can have different visions of the same reality. While one can model a relation as a link of composition, most models among the reviewed above treat it as a link of association.

The objective of our work consists in defining an algebra that allows the interrogation of documents in the context of an object-oriented approach. The basis of manipulation is the embedding of structures and the links between structures [Sch97]. This algebra permits to take into account the informational aspect as well as the structural aspect of documentary databases.

This paper is divided into 5 sections. The first one being this introduction, section 2 is devoted to the basic concepts of the proposed algebra. Section 3 is dedicated to the description of the algebraic operators and section 4 to examples of the algebraic operations. The last section consists in our conclusion and perspectives.

2. Formalization of a documentary database

In this part we define the basis of our algebra. Starting from a clean definition of what is a documentary database, we defined what is a subdatabase, the single type of our algebra and their operators which fit to the object-oriented ontology.

2.1. Documentary database

As usual in database theory, a documentary database is defined y a schema and an instance of this schema.

Definition 1: Documentary database schema

The documentary database schema can be defined by a finite acyclic connect graph. Formally, a schema $\Sigma$ of documentary database is defined by the triplet $(C, p, \lambda)$ where:

- $C = \{c_1, \ldots, c_n\}$ is a set of classes, where each class $c_i$ is characterized by some properties $p_i, p_{i+1}, \ldots, p_m$.
- $p : C \rightarrow 2^C$ is an hierarchy function which associates a set of classes to a given class. We note by $p^{-1}$ the inverse relation of the hierarchy function. $p^+$ the transitive hierarchy function of order n, $p^\ast$ the transitive hierarchy function and $p^\dagger$ the strict transitive hierarchy function which are defined as follows:

$\quad p^0(c) = c$
$\quad p^\dagger(c) = \bigcup \{ p^r(c) | r \geq 0 \}$
$\quad p^0(c) = c$
$\quad p^\dagger(c) = \bigcup \{ p^r(c) | r > 0 \}$
$\quad \lambda : C \times C \rightarrow 2^E$ is a Roll function which associates a set of rolls to a pair of classes. The schema is endowed by the following properties :

1) $\text{Root}(\Sigma) = \{ c \in C \mid \forall c \in C, \ c \notin \lambda(\text{Root}(\Sigma)) \} \cup \emptyset$ (due to the fact that the graph is acyclic and finite)
2) $(\forall c \in C)$, $(c \notin \lambda(p^\dagger(c)))$ (due to acyclicity)
3) $(\forall c \in \text{Root}(\Sigma))$, $(p^\dagger(c) \neq \emptyset)$ (this is the definition of the Root)
4) $(\forall c \in \text{Root}(\Sigma))$, $(p^\dagger(c) = p^\dagger(c))$ (due to the connection property)

Example:

Let $\Sigma = (C, p, \lambda)$ be the schema of the documentary database represented by fig. 1. The graph which corresponds to this schema is represented by fig. 2.

| 1 |
| 2 |
The documentary database schema can also be defined by a finite acyclic connected graph. We denoted it a $\Sigma$-graph.

Fig. 1. Example of a schema of a documentary database

$$C \equiv \{\text{Document (title, type, date-creation), Chapter (title, date-creation), Reference (reference), Section (title, date-creation), Content (text, type), Image (title, bitmap), Author (nom, name, address)}\}$$

$$p(\text{Document}) = \{\text{Author, Content, Chapter, Reference}\}$$
$$p(\text{Chapter}) = \{\text{Content, Section, Picture}\}$$
$$p(\text{Section}) = \{\text{Content, Picture}\}$$
$$p(\text{Reference}) = \{\text{Content, Picture}\}$$
$$p(\text{Publication}) = \{\text{Author (authors)}\}$$
$$p(\text{Content}) = \emptyset$$
$$p(\text{Picture}) = \emptyset$$
$$p(\text{Author}) = \emptyset$$

$$(1) \left\{ \text{Document, Author} \right\} = \{\text{main-author, co-authors}\}$$
$$(2) \left\{ \text{Document, Content} \right\} = \{\text{introduction, conclusion}\}$$
$$(3) \left\{ \text{Chapter, content} \right\} = \{\text{introduction, conclusion}\}$$
$$(4) \left\{ \text{Section, content} \right\} = \{\text{body}\}$$
$$(5) \left\{ \text{Section, Picture} \right\} = \{\text{pictures}\}$$
$$(6) \left\{ \text{Reference, Publication} \right\} = \{\text{publication}\}$$
$$(7) \left\{ \text{Chapter, Picture} \right\} = \{\text{pictures}\}$$
$$(8) \left\{ \text{Publication, Author} \right\} = \{\text{authors}\}$$

Fig. 2. Representation of the schema of the documentary database under the form of a connect acyclic graph

A documentary database is defined by a schema and by an instance of this schema.

**Definition 1**: Documentary database

A documentary database is defined by $\{\Sigma, \Omega, \mathcal{A}\}$ where $\Sigma / \mathcal{C}, p, \lambda)$ is a schema of the database, $\mathcal{O}$ is a set of objects and $\mathcal{A}$ a set of inter-objects links $\mathcal{A} \subseteq \mathcal{O} \times \mathcal{O}$. If we note by $\tau$ the total function (\tau: \mathcal{O} \rightarrow \mathcal{C}) which returns for every object its class, then $\{\Sigma, \Omega, \mathcal{A}\}$ must verify the following constraint:

- $\forall (a_i, a_j) \in \mathcal{A}, \forall (a_i) \in \mathcal{C}$

**2.2. Documentary subdatabase**

A user who wants to formulate queries on a database, is rarely interested by the whole entities in the database. In most cases, he wants to formulate queries on a portion of this database. For example, a user would like to formulate queries only on the Author class and Document class of our illustration.
We defined a subdatabase of a database D as a database which is merged into the database D.

**Definition 3 : Subschema**

A subschema \( \psi = (C_\psi, \rho_\psi, \lambda_\psi) \) of a schema \( \Sigma = (C_\Sigma, \rho_\Sigma, \lambda_\Sigma) \) is included in the \( \Sigma \) graph, such that:
- \( C_\psi \subseteq C_\Sigma \)
- \( \forall e \in C_\psi \; \rho_\psi(e) \subseteq \rho_\Sigma(e) \)
- \( \forall \psi, c_\psi \in C_\psi \; \lambda_\psi(e_c, c_\psi) \subseteq \lambda_\Sigma(e, c_\psi) \)

**Remarks:**
- A schema is a subschema of itself.
- A class of a schema is a subschema of the schema.

Example:
Fig. 4 represents the graph of the subschema \( \psi = (C_\psi, \rho_\psi, \lambda_\psi) \) which is defined as:
\[
\psi = \{ \text{Document (title, type, creation-date), Chapter (title, creation-date), Reference (reference)} \}
\]
\[
\rho_\psi(\text{Document}) = \{ \text{Chapter, Reference} \}
\]
\[
\rho_\psi(\text{Reference}) = \emptyset
\]

![Fig. 4 : graphical representation of a subschema](image)

**Definition 4 : An instance of a subschema**

An instance of a subschema \( \psi = (C_\psi, \rho_\psi, \lambda_\psi) \) is a set of objects and a set of links between objects which respectively correspond to the classes in the subschema and to the links between classes. Formally, an instance is defined by the pair:
\[
(O, A)
\]
where \( O \) is a set of objects and \( A \) a set of inter-objects links \( A \subseteq O \times O \) with the following constraint:
- \( \forall (x, y) \in A \; \forall (y, z) \in A \; x \neq y \neq z \)

Example:
Fig. 5 gives a graphical representations of an instance of the subschema that is represented in fig. 4.

![Fig. 5 : An example of an instance of a subschema](image)

**Definition 5 : Subdatabase**

A subdatabase is defined by a subschema and by an instance of this subschema. Formally, a subdatabase is defined by the triple:
\[
(\psi, O, A)
\]

Example :
Definition 8: Distance between two subschemas

Given two subschemas \( \psi_1 = (C_{\psi_1}, P_{\psi_1}, \Lambda_{\psi_1}) \) and \( \psi_2 = (C_{\psi_2}, P_{\psi_2}, \Lambda_{\psi_2}) \), the distance between \( \psi_1 \) and \( \psi_2 \), denoted by \( D(\psi_1, \psi_2) \), is defined as:

\[
D(\psi_1, \psi_2) = \min \left( \text{Len}(P_1) / P \in \text{SP}(\psi_1, \psi_2) \right)
\]

Example:
The following is a distance between subschemas \( \psi_1 \) and \( \psi_2 \) shown in figure 6:

\[
D(\psi_1, \psi_2) = \min (\text{Len(Document.Reference)}, \text{Len (Document.Author.Publication)}) = 2
\]

Definition 9: The upper limit of two subschemas

Given two subschemas \( \psi_1 = (C_{\psi_1}, P_{\psi_1}, \Lambda_{\psi_1}) \) and \( \psi_2 = (C_{\psi_2}, P_{\psi_2}, \Lambda_{\psi_2}) \), the upper limit \( \psi_1 \vee \psi_2 \) is the least subschema containing \( \psi_1 \) and \( \psi_2 \). The upper limit can be defined as the union of \( \psi_1 \), \( \psi_2 \) and all the paths between \( \psi_1 \) and \( \psi_2 \) that have their lengths equal to \( D(\psi_1, \psi_2) \). Let us consider the set \( S \) of all the paths between \( \psi_1 \) and \( \psi_2 \) that have their lengths equal to \( D(\psi_1, \psi_2) \). The set \( S \) is defined as:

\[
S = \{ \text{Path}_{\psi_1, \psi_2} | \text{Len}(\text{Path}_{\psi_1, \psi_2}) = D(\psi_1, \psi_2) \}
\]

Formally, the upper limit of two subschemas \( \psi_1 = (C_{\psi_1}, P_{\psi_1}, \Lambda_{\psi_1}) \) and \( \psi_2 = (C_{\psi_2}, P_{\psi_2}, \Lambda_{\psi_2}) \), is defined as:

\[
\psi_1 \vee \psi_2 = (C_{\psi_1}, P_{\psi_1} \cup P_{\psi_2}, \Lambda_{\psi_1} \cup \Lambda_{\psi_2} \cup \text{Path}_{\psi_1, \psi_2})
\]

where:

- \( C_{\psi_1, \psi_2} = C_{\psi_1} \cup C_{\psi_2} \)
- \( \text{Path}_{\psi_1, \psi_2} = \text{Path}_{\psi_1} \cup \text{Path}_{\psi_2} \)

Example:
An example of an upper limit of two subschemas is shown in fig. 8.

---

5. Algebraic operators

We have shown that the set of subschema of a database is closed under \( \sqcap \) and \( \vee \). We now prove that this set is also closed under algebraic operators: select, project, associate, union, intersection, difference, join, grouping, split, distribution.
The algebraic operators have as operands a subdatabase and produce a new subdatabase. In this algebra, the closure property is maintained, because the results produced by the queries are structured in the same manner as the operands. Also, this algebra allows to produce a new class of objects that enriches the existing database.

The algebraic operators will be formally defined in the appendix. The examples used to explain these operators will make use of the database shown in fig. 1 and 3. We extend algebraic operators to subdatabases.

3.1. Select (∀)

The Select is a unary operator, which operates on a subdatabase to produce a new subdatabase in which objects and inter-objects links satisfy a specified predicate. The resultant subdatabase has the same schema as the operand subdatabase, and the instances of the result is a subset of the instances of the operands which satisfy the predicate. The select operation is denoted as \( \forall (X) [P] \) where \( X \) is an operand subdatabase and \( P \) is a predicate. The predicate \( P \) is a logical expression that is evaluated to true or false. This logical expression is composed by terms interconnected by logic operators (and, or, not). Each term can be, either a condition on an attribute value, or a condition on an inter-object link. The condition on the attribute values have the following form:

- \( \theta \) constant, or \( a \theta b \)

where \( a \) and \( b \) represent attributes.

- If \( a, b \) and constant are integer, then \( \theta \) can be: =, \(!=, <, >, \leq, \geq\)
- If \( a, b \) and constant are Boolean, then \( \theta \) can be: =, \(!=\)
- If \( a, b \) and constant are string, then \( \theta \) can be: =, \(!=, <, >\)
- If \( a \) is a text and constant is a string, then \( \theta \) can be: contains, notcontains

To express the conditions on the cardinalities of links, we use the symbol \( \theta \) link. The conditions on the cardinalities of links have the following form: \( \delta \theta \) link, where \( \delta \) and \( \beta \) are classes, \( \theta \) is a comparison operator that can be: =, \(!=, <, >, \leq, \geq\) and \( \delta \) is a numeric constant. For example, the predicate \( "Document" \& \& \"Author \&\& \"2\" \) means that the object of the Document class must be linked to only two objects of the Author class.

Example:

- Find the documents which are written by two authors only, whose titles contain the word "computer" and which are composed of at least two chapters.

The algebraic expression of this query is:

\[
X := \forall (\text{Document, Author, Chapter, \langle Document, Author, Chapter \rangle}) (\text{Document.title contains "computer" and Document.\#Chapter = 2})
\]

If we suppose that the values of the title attribute (cf fig. 2) are:

- d.title = "Computer and medicine"
- d.title = "Data and Computer"
- d.title = "Data Processing"
- d.title = "Computer Science"

The result of the select operation is shown in fig. 11.

3.2. Projection (∧)

The project is a unary operator, which operates on a subdatabase to produce a new subdatabase reduced to a subset of classes, a subset of inter-class links, and a subset of attributes. The project operation is denoted \( \pi (X) [E (P_1, P_2, ..., P_n)] \) where \( X \) is the subdatabase operand and the set \( \{P_1, P_2, ..., P_n\} \) represents the different classes of the projected subdatabase. The set \( \{P_1, P_2, ..., P_n\} \) represents the properties of the projection of the \( c_i \) class.

The project operation is valid only if the subclass \( \forall (X) [P_1, P_2, ..., P_n] \) is such as:

- \( C_i \subset C_X \)
- \( \forall (X) [P_1, P_2, ..., P_n] \subset (\forall (X) [P_1] \cap C_Y) \)

Example:

An example of projection operation is shown in fig. 12. The expression of this projection operation is:

\[
Y := \pi (X) [\langle \text{Document.title, Author.first-name, name} \rangle]
\]

3.3. Associate (⊗)

The associate operator is a binary operator which constructs a new subdatabase by concatenating two subdatabases. The subschema of the result subdatabase is a concatenation of the subschemas of operands subdatabases through some inter-classes links. The instance of result subdatabase is constituted by the objects of the operands which are connected together. Since two subdatabase may have more than one link, it is necessary to specify through which links the concatenation is made. The associate operation between subdatabase \( X \) and \( Y \) through the links \( L \) is denoted \( X \otimes L \ Y \), where \( L = \forall (c_i, c_j) \) is included in the set \( C_X \times C_Y \).

Example:

Let \( X \) and \( Y \) be the subdatabases represented by fig. 13, an example of associate operation is:

\[
Z := X \otimes \forall (\text{Section, Picture}) \ Y
\]

The result subdatabase of this query is shown in fig. 13. In this example, the object \( i \) of the Picture class is not represented in the result because it is not linked to any object of the set \( \{s_1, s_2, s_3, s_4, s_5, s_6, s_7\} \) (cf. fig. 2). For the same reason, the objects \( \{c_1, c_2, c_3, c_4\} \) of the Chapter class are also omitted from the result because they are not linked to any of the objects of the Section class.
3.4. Union, Intersection and Difference (∪, ∩, −)

The union operator, as a binary operator which constructs a new subdatabase by combining two subdatabases. The union operation is denoted X ∪ Y.

The intersection operator, as the union operator, is a binary operator which constructs a new subdatabase by combining two subdatabases. The intersection operation is denoted X ∩ Y.

The difference operator, as the union and intersection operators, is a binary operator which constructs a new subdatabase by subtracting from a subdatabase X of an other subdatabase Y. The difference operation is denoted X − Y.

Example:
Examples of union, intersection and difference operations are shown in fig. 14.

3.5. Join

Unlike the relational data model, in object-oriented models many relationships between objects can be represented within the objects themselves. Hence the join operations are used less frequently in the object algebra than in the relational algebra. However, the existing objects in a database may not explicitly reflect all relationships required by the queries. An explicit join is still needed to handle the cases when the relationship being queried upon is not defined within the object classes. Such relationship is called value-based relationship in contrast with those specified explicitly in the object model. For example, suppose we have a layer document class and a technical document class. A query «Find the layer document which has the same key words as the technical document» requires a value-based join between the classes layer document and technical document through the attribute keywords.

The join condition links the properties of only one class of X and the properties of only one class in Y. The subschema of the result subdatabase is defined as the concatenation of the subschemas of the subdatabases operands through a new class c. This new class is defined as a set of tuples. The tuple is made of two attributes, the first attribute is a set of references to the class c1 of X, the second attribute is a set of references to a class c2 of Y. The graphical structure of class c is shown in fig 15. The instance of the result subdatabase is constituted of a portion of X that satisfies the join condition, a portion of Y that satisfies the join condition, and the union of the instances of
operands which are extended to the subschema of the result subdatabase. The Join operation is denoted by $X \bowtie [P] Y$ where $X$ and $Y$ are the subdatabases operands and $P$ represents the join predicate.

**Example:**

Let $X$ and $Y$ be the subdatabases represented by the fig. 16. An example of join operation is:

$$X \bowtie [\text{Chapter-date-creation} = \text{Part-date-creation}] Y$$

if, we suppose that the values of the attributes $\text{date-creation}$ of the $\text{reference-guide}$ class and of the $\text{user-guide}$ class are:

- $c_c, \text{creation-date} = "26-03-82", c_r, \text{creation-date} = "26-03-82", c_e, \text{creation-date} = "31-03-82", p_p, \text{creation-date} = "26-03-82", p_f, \text{creation-date} = "26-03-82", p_e, \text{creation-date} = "03-07-82", p_r, \text{creation-date} = "03-07-82"$

The result of the join operation is shown in fig. 16. The subschema of the result subdatabase is composed of the subschema of $X$ and of $Y$, and of a new $\text{Chapter-Part}$ class which refer to $\text{Chapter}$ class and $\text{Part}$ class. The $\text{Chapter-Part}$ class is composed of two objects $f_i$ and $f_j$. The object $f_i$ is linked to the objects $\{c_i, c_j\}$ of the $\text{Chapter}$ class and it is also linked to the objects $\{p_i, p_j\}$ because the two sets of objects have the same value of the $\text{date-creation}$ attribute and the join condition is fulfilled. For the same reason, the object $f_j$ is linked to the objects $\{p_i, p_j\}$. However, the object $c_o$ of Chapter class is not present in the result, because it doesn’t fulfill the join condition. Consequently, the objects $\{c_i, c_j\}$ of the $\text{Selection}$ class are also omitted from the result.

### 3.6 Grouping ($\gamma$)

The grouping is an unary operator. This operator is also used in NST-Algebra [Gut89]. It is used to transform an association relationship to an aggregation relationship. The grouping operation is defined as $\gamma(X)[c_1, c_2, a]$ where $X$ is an operand subdatabase and $\{c_1, c_2, a\}$ is the association relationship to be transformed into an aggregation relationship.

**Example:**

An example of a grouping operation is shown in fig. 17.

### 3.7 Split ($\delta$)

The split operation is the inverse of the grouping operation. Indeed, the split operation is used to transform an aggregation relationship to an association relationship. The split operation is defined as $\delta(X)[c_1, c_2, a]$ where $X$ is an operand subdatabase and $\{c_1, c_2, a\}$ is the aggregation relationship to be transformed into an association relationship.

**Example:**

An example of the split operation is shown in fig 18.
3.8 Distribution (Ω)

The distribution operator is used to restructure a subdatabase. This restructuration is made by distributing a property of class $c_3$ to a class $c_2$. The distribution is valid, if there exists a link between $c_2$ and $c_3$. The distribution operation is defined as $\delta(X)\{c_3, c_1, p\}$ where $X$ is a subdatabase operand, $p$ is the property of $c_1$ to be distributed to class $c_2$.

Example:

An example of the distribution operation is shown in figure 19.

4. Examples of expressions of algebraic queries

A storage and a query system of structured documents using the OODMBS technology was proposed in [Ch94]. This system was developed in the Varso project of INRIA. It is based on an extension of the model and query language of O2 OODMBS [Mes90]. In order to request documents by their structure and content, O2SQL was extended to:

- take into account new functions of modeling like the ordered n-tuples and unions of types
- include predicates of textual retrieval like contains.

We will present examples of queries to compare our algebraic approach with the O2SQL query language.

Example 1:

«Find the documents containing the word Computer in the title and which are constituted of two chapters.»

The request is written as follows in the O2SQL query language:

\begin{align*}
\text{Select} & \quad s \\
\text{From} & \quad x \text{ in Document} \\
\text{Where} & \quad x.title \text{ contains «computer» and count (x.chapter)} = 2
\end{align*}

The algebraic expressions of the request are:

\begin{align*}
R_1 := & \forall [\text{Document, Chapter}] ((\text{Document.title contains «computer»} \\
& \quad \land \quad \text{count (Document.chapter)} = 2)) \\
R_2 := & \forall [R_1] (\text{Document}) \\
R_3 := & \forall [\text{Document, Chapter}] (\text{Document.title contains «computer»} \\
& \quad \land \quad \text{count (Document.chapter)} = 2)
\end{align*}

Example 2:

«Find the documents that are only constituted of two chapters with the following condition: each chapter belonging to three documents.»

The algebraic expression of the query is:

\begin{align*}
R_1 := & \forall [\text{Document, Chapter}] (\text{Document.chapter} = 2) \\
R_2 := & \forall [R_1] (\text{Document}) \\
R_3 := & \forall [R_1] (\text{Document})
\end{align*}

To our knowledge, this query cannot be expressed in O2SQL. Indeed, such a query expresses a condition on the structure starting from the document to chapters and inversely, from chapters to the document.

5. Conclusion

We have presented in this paper an algebra for structured documents in an object-oriented approach. We have defined the concept of subdatabase on which is based the definition of algebraic operators. Indeed, the operators and the query result are subdatabases, as a consequence, the closing condition and the operators orthogonality conditions are respected because the operands and the results of the query operators are structured in the same way. The operators can be regrouped into two categories. A first category that deals with information retrieval operators (select, project, associate, union, intersection, difference and join operators). A second category that deals with information restructuring operators (grouping, split, distribution operators).

Appendix

The formal definition of the algebraic operators are given below:

- \textbf{Select}

Let us consider a subdatabase $X = (\psi, X, O_X, A_X) \land (\psi, X = (C, \rho, \lambda, \chi))$. The select operation is defined as:

\begin{align*}
Y & = \sigma(X / P) = \psi_X \land (\psi_X = (C, \rho, \lambda, \chi)) \\
& = \forall \quad C_Y = C_X \\
& \quad \land \quad \psi_Y = \psi_X = (C_Y, \rho_Y, \lambda_Y) \\
& \quad \land \quad \Pi_Y = \psi_Y = \psi_X = (C_Y, \rho_Y, \lambda_Y)
\end{align*}

- \textbf{Project}

Let us consider a subdatabase $X = (\psi, X, O_X, A_X) \land (\psi, X = (C, \rho, \lambda, \chi))$. The project operation is defined as:

\begin{align*}
I & = \pi(X / \{c_1, c_2, \ldots, c_n\}) = (\psi_X, O_X, A_X) \land (\psi_X = (C, \rho, \lambda, \chi)) \\
& = \forall \quad C_Y = \{c_1, c_2, \ldots, c_n\} \\
& \quad \land \quad \psi_Y = (\psi_X, O_X, A_X) \\
& \quad \land \quad \Pi_Y = (\psi_X, O_X, A_X)
\end{align*}

- \textbf{Associate}

...
Let us consider a subdatabase $X=(\psi_X, O_X, A_X)$ ($\psi_X=(C_X, p_X, \lambda_X)$) and a database $D=(\mathbb{Z}, \mathbb{A})$. The association operation is defined as:

$$Z = X \bowtie D = \{ (\psi_X, O_X, A_X) \}$$

where $\psi_Z = (C_Z, p_Z, \lambda_Z)$ such as:

- $C_Z = C_X \cup C_Y$
- $O_Z = O_X \cup O_Y$
- $A_Z = A_X \cup A_Y$

- Union

Let us consider a subdatabase $X=(\psi_X, O_X, A_X)$ ($\psi_X=(C_X, p_X, \lambda_X)$) and $Y=(\psi_Y, O_Y, A_Y)$ ($\psi_Y=(C_Y, p_Y, \lambda_Y)$). The union operation is defined as:

$$Z = X \cup Y = \{ (\psi_X, O_X, A_X) \}$$

where $\psi_Z = (C_X, p_X, \lambda_X)$ such that:

- $\psi_Z = \psi_X$
- $\psi_Z = \psi_Y$
- $O_Z = O_X \cup O_Y$
- $A_Z = A_X \cup A_Y$

- Intersection

Let us consider a subdatabase $X=(\psi_X, O_X, A_X)$ ($\psi_X=(C_X, p_X, \lambda_X)$) and $Y=(\psi_Y, O_Y, A_Y)$ ($\psi_Y=(C_Y, p_Y, \lambda_Y)$). The intersection operation is defined as:

$$Z = X \cap Y = \{ (\psi_X, O_X, A_X) \}$$

where $\psi_Z = (C_X, p_X, \lambda_X)$ such as:

- $\psi_Z = \psi_X$
- $\psi_Z = \psi_Y$
- $O_Z = O_X \cap O_Y$
- $A_Z = A_X \cap A_Y$

- Difference

Let us consider a subdatabase $X=(\psi_X, O_X, A_X)$ ($\psi_X=(C_X, p_X, \lambda_X)$) and $Y=(\psi_Y, O_Y, A_Y)$ ($\psi_Y=(C_Y, p_Y, \lambda_Y)$). The difference operation is defined as:

$$Z = X - Y = \{ (\psi_X, O_X, A_X) \}$$

where $\psi_Z = (C_X, p_X, \lambda_X)$ such as:

- $\psi_Z = \psi_X$
- $O_Z = O_X \cap O_Y$
- $A_Z = A_X \cap A_Y$

- Join

Let us consider a subdatabase $X=(\psi_X, O_X, A_X)$ ($\psi_X=(C_X, p_X, \lambda_X)$) and $Y=(\psi_Y, O_Y, A_Y)$ ($\psi_Y=(C_Y, p_Y, \lambda_Y)$). The join operation is defined as:

$$Z = X \bowtie [\mathbb{R}(c, d)] \{ (\psi_X, O_X, A_X) \}$$

where $\psi_Z = (C_Z, p_Z, \lambda_Z)$ such as:


