Creating Structure in Unstructured Data

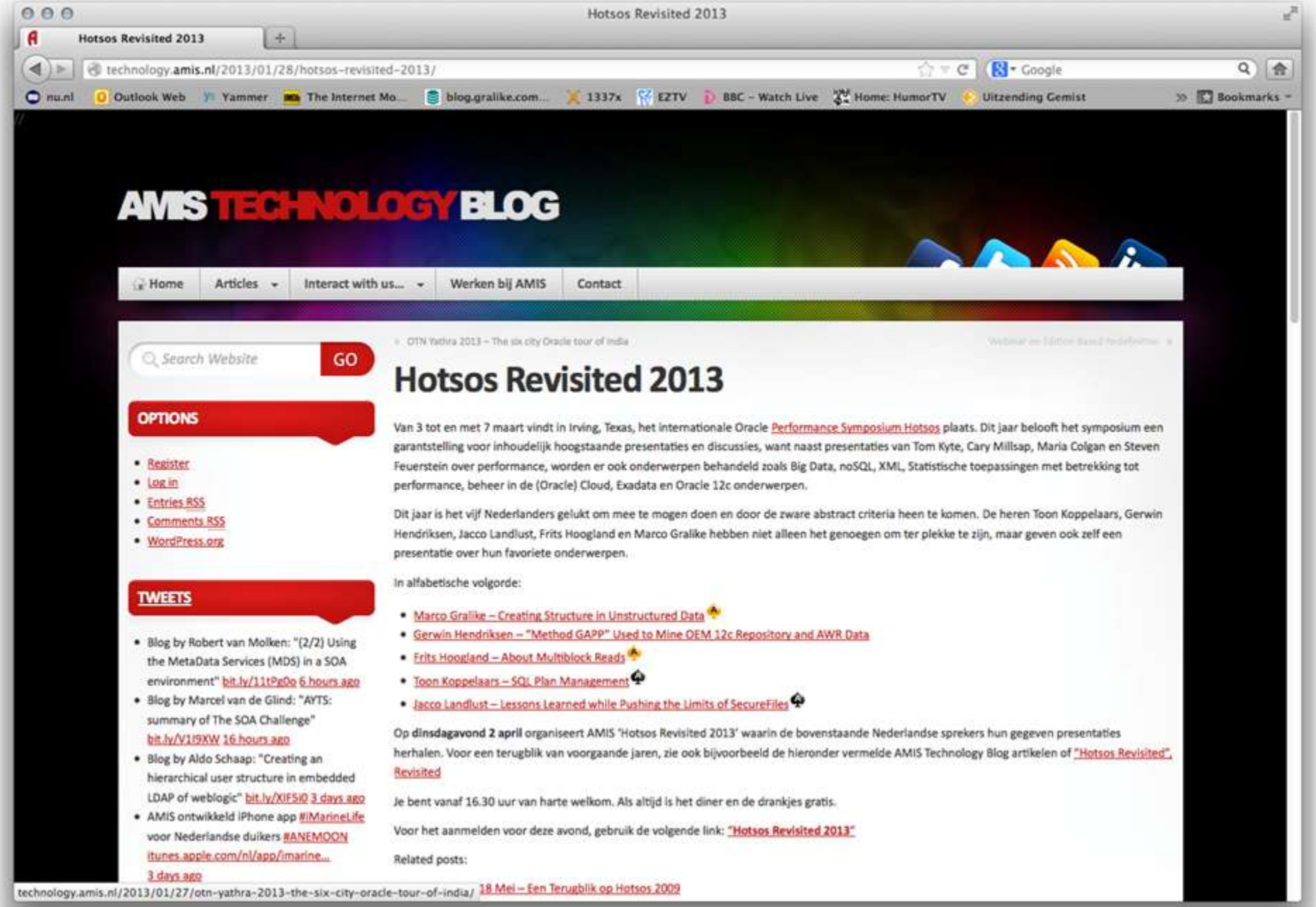
What is possible, today…?

Marco Gralike

ORACLE

AMIS

OakTable.net


Hotsos Revisited 2013
Hotsos Revisited 2013
technology.amis.nl/2013/01/28/hotsos-revisited-2013/
Google
nu.nl
Outlook Web
Yammer
The Internet Mo...
blog.gralike.com...
1337x
EZTV
BBC - Watch Live
Home: HumorTV
Uitzending Gemist
Bookmarks
AMIS TECHNOLOGY BLOG
Home
Articles
Interact with us...
Werken bij AMIS
Contact
Search Website
GO
OPTIONS
Register
Log in
Entries RSS
Comments RSS
WordPress.org
TWEETS
Blog by Robert van Molken: "(2/2) Using the MetaData Services (MDS) in a SOA environment" bit.ly/11tPg0o 6 hours ago
Blog by Marcel van de Glind: "AYTS: summary of The SOA Challenge" bit.ly/V1I9XW 16 hours ago
Blog by Aldo Schaap: "Creating an hierarchical user structure in embedded LDAP of weblogic" bit.ly/XIFSi0 3 days ago
AMIS ontwikkeld iPhone app #iMarineLife voor Nederlandse duikers #ANEMOON itunes.apple.com/nl/app/imarine... 3 days ago
OTN Yathra 2013 – The six city Oracle tour of India
Hotsos Revisited 2013
Van 3 tot en met 7 maart vindt in Irving, Texas, het internationale Oracle Performance Symposium Hotsos plaats. Dit jaar belooft het symposium een garantstelling voor inhoudelijk hoogstaande presentaties en discussies, want naast presentaties van Tom Kyte, Cary Millsap, Maria Colgan en Steven Feuerstein over performance, worden er ook onderwerpen behandeld zoals Big Data, noSQL, XML, Statistische toepassingen met betrekking tot performance, beheer in de (Oracle) Cloud, Exadata en Oracle 12c onderwerpen.
Dit jaar is het vijf Nederlanders gelukt om mee te mogen doen en door de zware abstract criteria heen te komen. De heren Toon Koppelaars, Gerwin Hendriksen, Jacco Landlust, Frits Hoogland en Marco Gralike hebben niet alleen het genoegen om ter plekke te zijn, maar geven ook zelf een presentatie over hun favoriete onderwerpen.
In alfabetische volgorde:
Marco Gralike – Creating Structure in Unstructured Data
Gerwin Hendriksen – "Method GAPP" Used to Mine OEM 12c Repository and AWR Data
Frits Hoogland – About Multiblock Reads
Toon Koppelaars – SQL Plan Management
Jacco Landlust – Lessons Learned while Pushing the Limits of SecureFiles
Op dinsdagavond 2 april organiseert AMIS 'Hotsos Revisited 2013' waarin de bovenstaande Nederlandse sprekers hun gegeven presentaties herhalen. Voor een terugblik van voorgaande jaren, zie ook bijvoorbeeld de hieronder vermelde AMIS Technology Blog artikelen of "Hotsos Revisited", Revisited
Je bent vanaf 16.30 uur van harte welkom. Als altijd is het diner en de drankjes gratis.
Voor het aanmelden voor deze avond, gebruik de volgende link: "Hotsos Revisited 2013"
Related posts:
18 Mei – Een Terugblik op Hotsos 2009
technology.amis.nl/2013/01/27/otn-yathra-2013-the-six-city-oracle-tour-of-india/


www.xmldb.nl » About Oracle, XMLDB and other interests...
www.xmldb.nl » About Oracle, X...
www.liberidu.com/blog/
Google
nu.nl
Outlook Web
Yammer
The Internet Mo...
blog.gralike.com...
1337x
EZTV
BBC - Watch Live
Home: HumorTV
Uitzending Gemist
Bookmarks
www.xmldb.nl
About Oracle, XMLDB and other interests...
About
Papers
OTN
Sites
XFiles
XML Content
Old Stuff
Music
[Menu]
Search
JAN 28 2013
Hotsos Revisited 2013
11gR1, 11gR2, 12cR1, Events, Oracle, Performance, RDBMS, Unstructured Data
hotsos
Van 3 tot en met 7 maart vindt in Irving, Texas, het internationale Oracle Performance Symposium Hotsos plaats. Dit jaar belooft het symposium een garantstelling voor inhoudelijk hoogstaande presentaties en discussies, want naast presentaties van Tom Kyte, Cary Millsap, Maria Colgan en Steven Feuerstein over performance, worden er ook onderwerpen behandeld zoals Big Data, noSQL, ...
Continue reading »
Tags: Frits Hoogland, gerwin hendriksen, hotsos, hotsos 2013, Jacco Landlust, Marco Gralike, Toon Koppelaars
Leave comment
JAN 21 2013
Basisregistraties Adressen en Gebouwen – Het importeren van Kadaster BAG data in een Oracle Database
Binary-, CLOB-, Object Relational Storage, Howto, Oracle, XMLDB
Vorig jaar heb ik behoorlijk wat vragen gekregen over of er een tool was, of een methodiek, om BAG data van het Nederlandse Kadaster in een Oracle database te krijgen voor allerlei doeleinden. BAG data (Basisregistraties Adressen en Gebouwen) wordt, zo ver ik weet, onder andere uitgeleverd door het Kadaster in XML bestanden waarin alle ...
AMIS Company Profile
AMIS
Click here for more info
Appearances 2012/2013
ODTUG Webcast (19 Jan)
RMOUG (February 14-16)
2 Day Masterclass Slovenia (March 27-28)
Oracle Open World (Sep 30-Oct 04)
BGOUG (November 16-18)
Hotsos (March 3-7)
10 Most Popular (7 days)
Solving VMware network problems on Linux VMware guests 198 view(s) | posted on September 29, 2006
Windows "SC" command – Handling Windows

“Big Data” = XML ?

Challenges are!

Ahum, the problems are!

WikiPedia

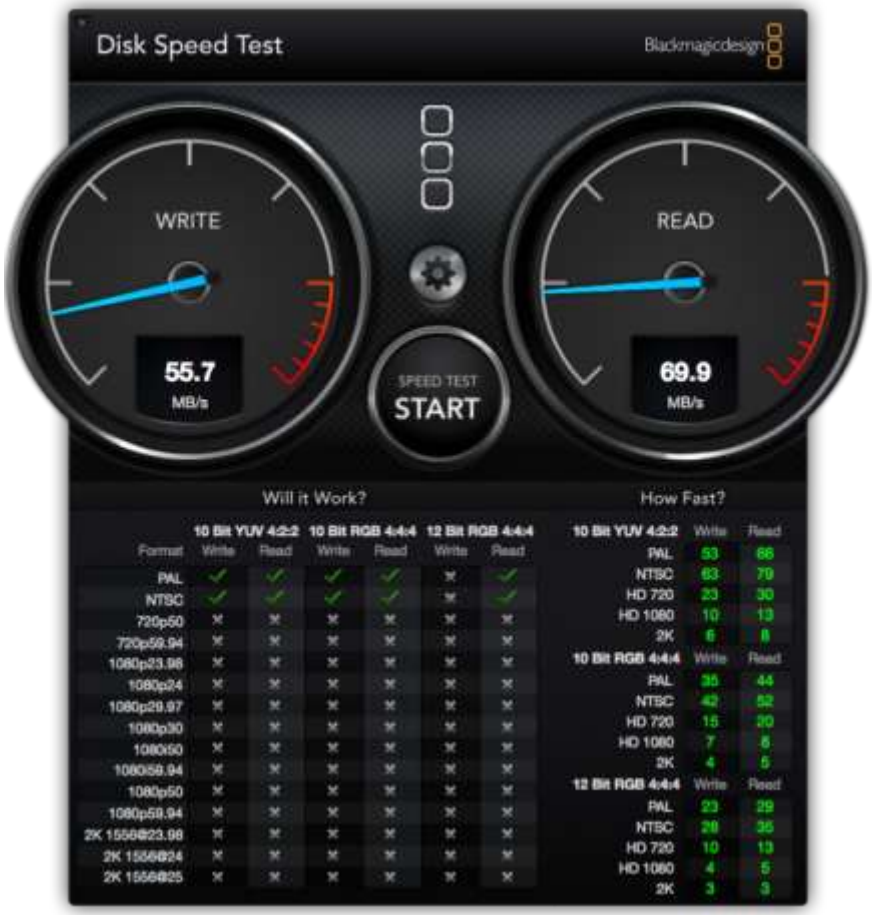
- One string of XML data with structured and unstructured data sections
- Language: English
- Size : 42,15 GB
- Pages : 12.961.997
- Date : 21 Dec 2012

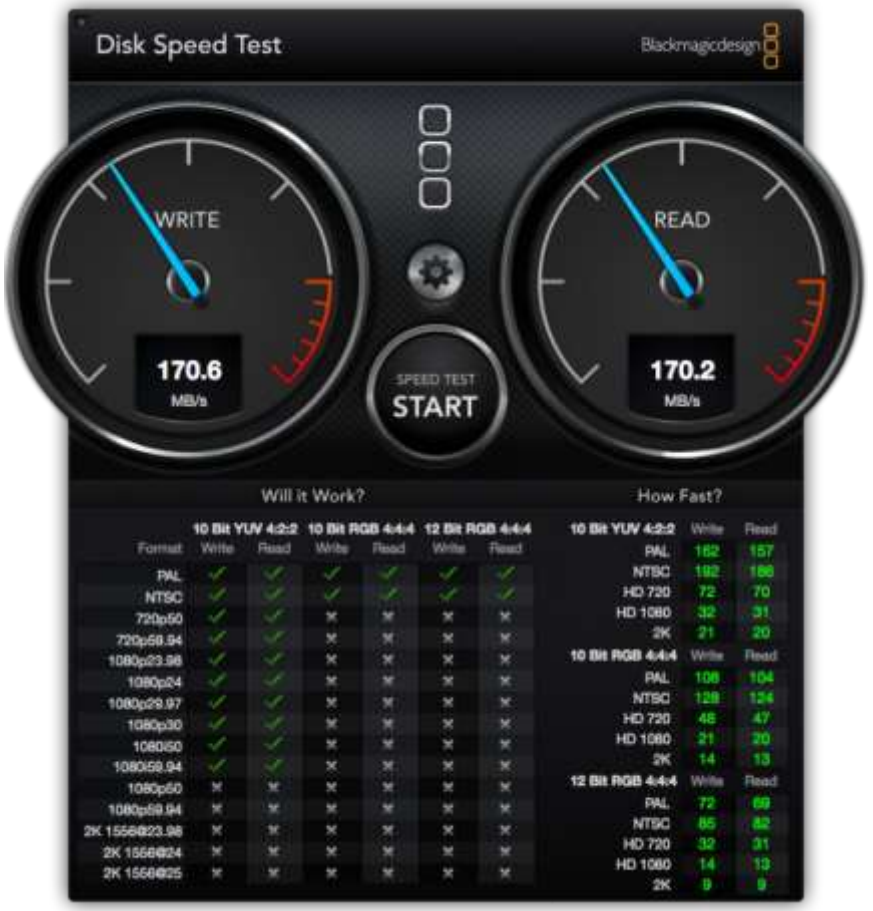
Adventures into the unknown…?

Setup

- VirtualBox VM
  - OEL 5U8 (64)
  - 8 GB RAM
- LaCie Little Big Disk
  - RAID 0
  - **Thunderbolt**
- Database
  - SGA 4GB
  - PGA 2GB

My new LaCie LBD is really fast - ☺

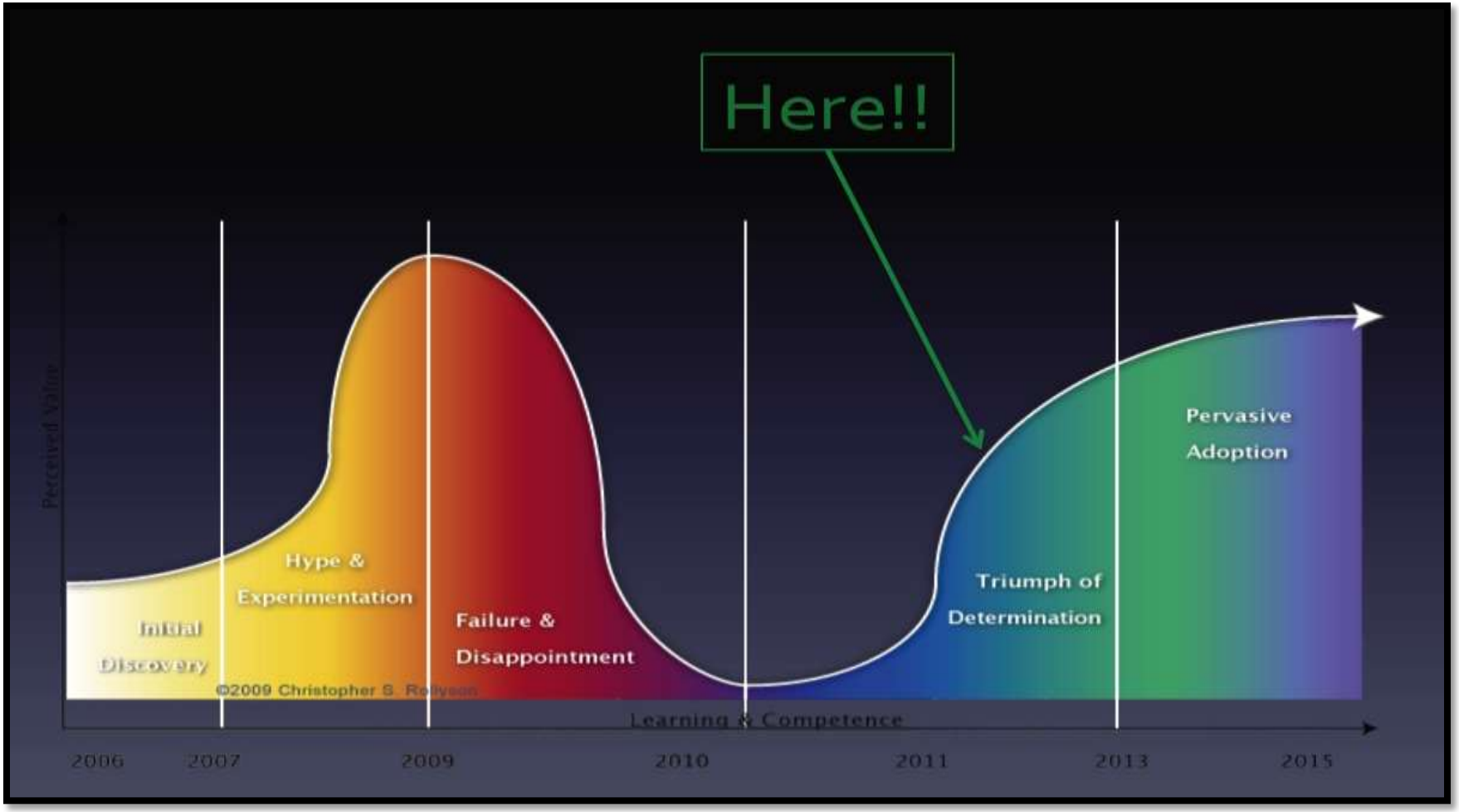




Defeat?! - 1.000.000 pages only

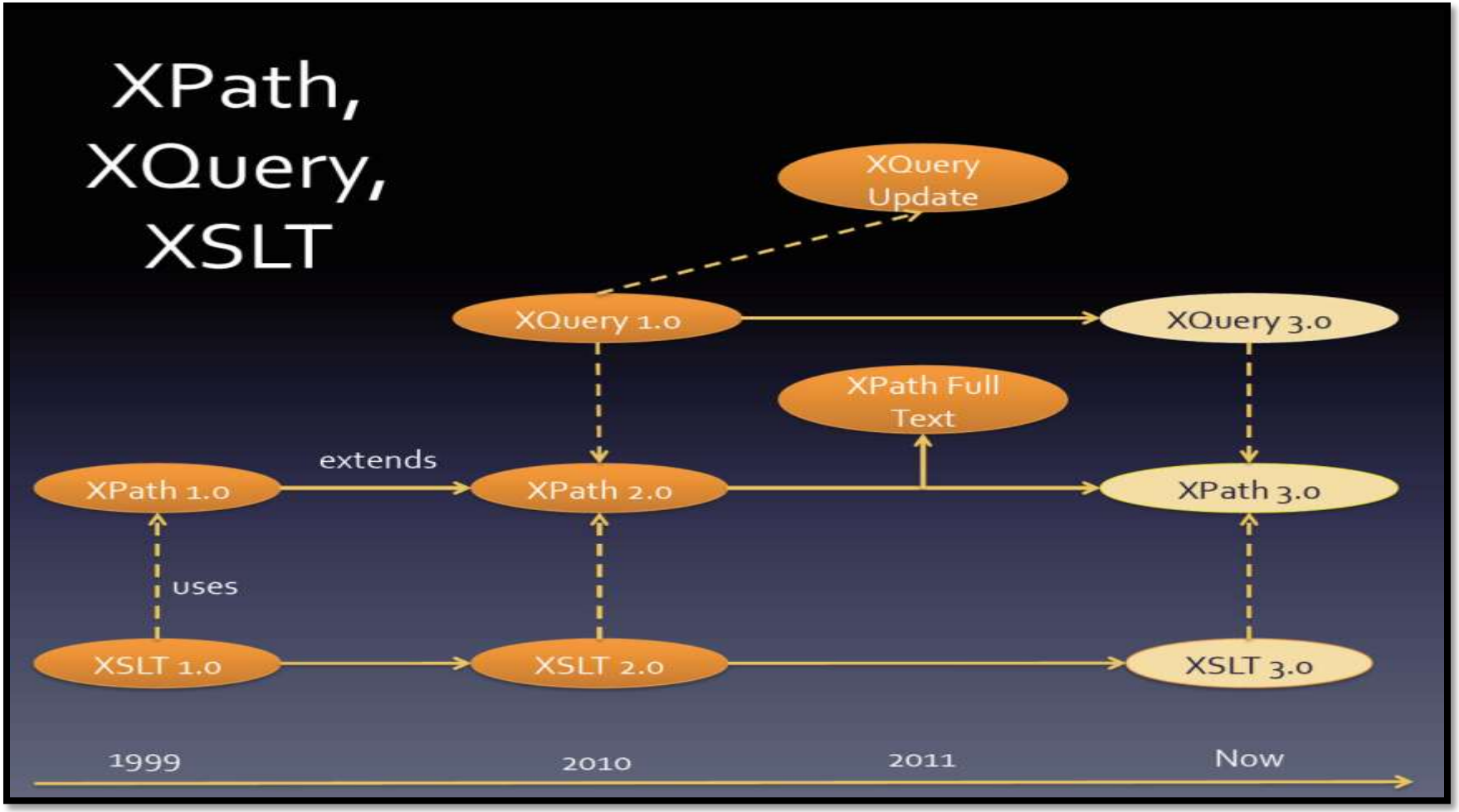
Status of Technology used

XML - Where are we…?



Gartner

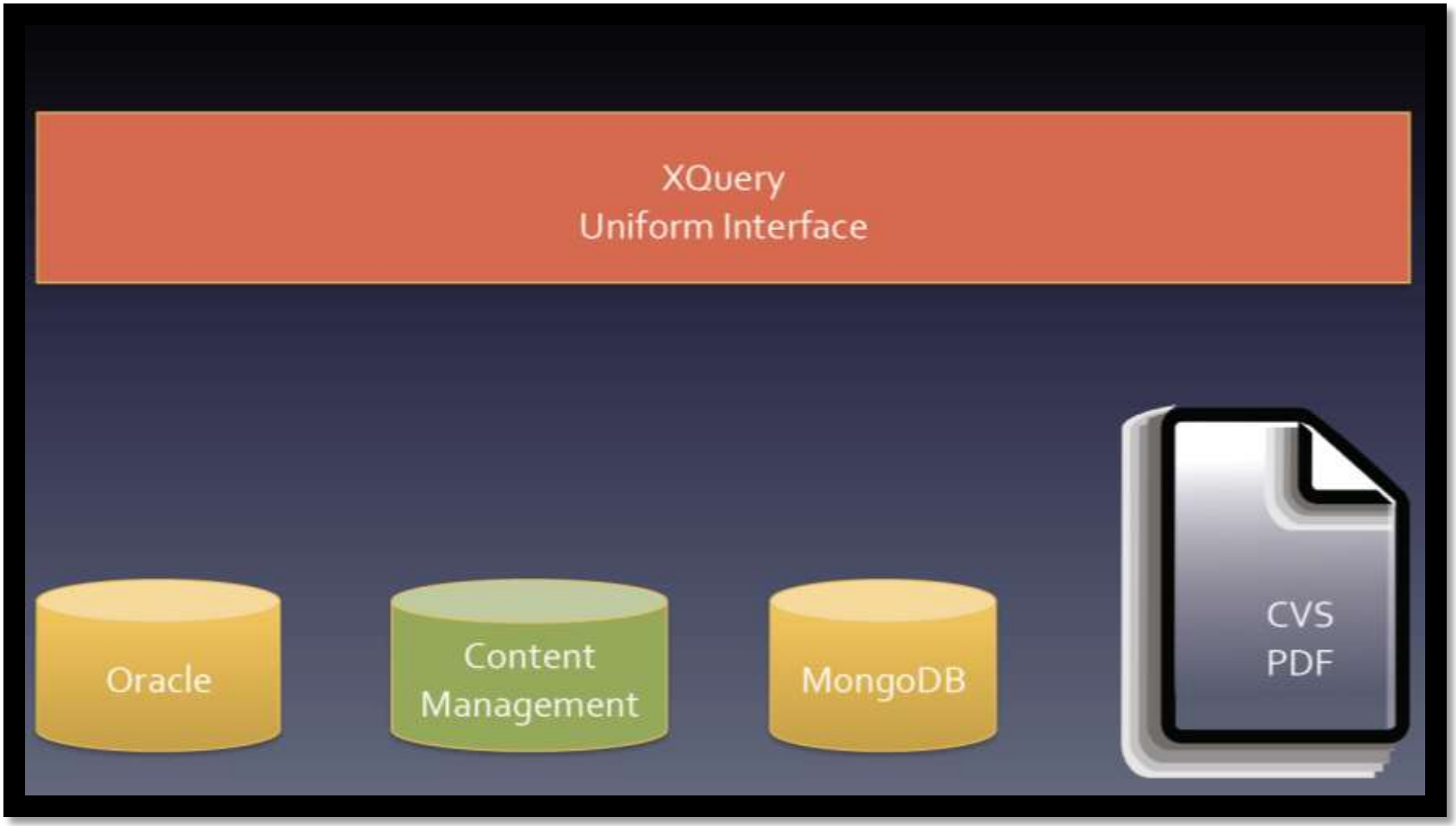
Achieved…?


XPath,
XQuery,
XSLT
XQuery Update
XQuery 1.0
XQuery 3.0
XPath Full Text
extends
XPath 1.0
XPath 2.0
XPath 3.0
uses
XSLT 1.0
XSLT 2.0
XSLT 3.0
1999
2010
2011
Now

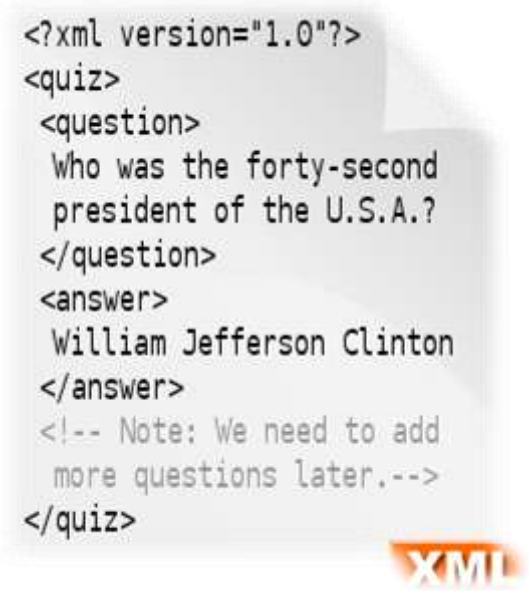
On the Horizon!

- JSoniq
- Zorba

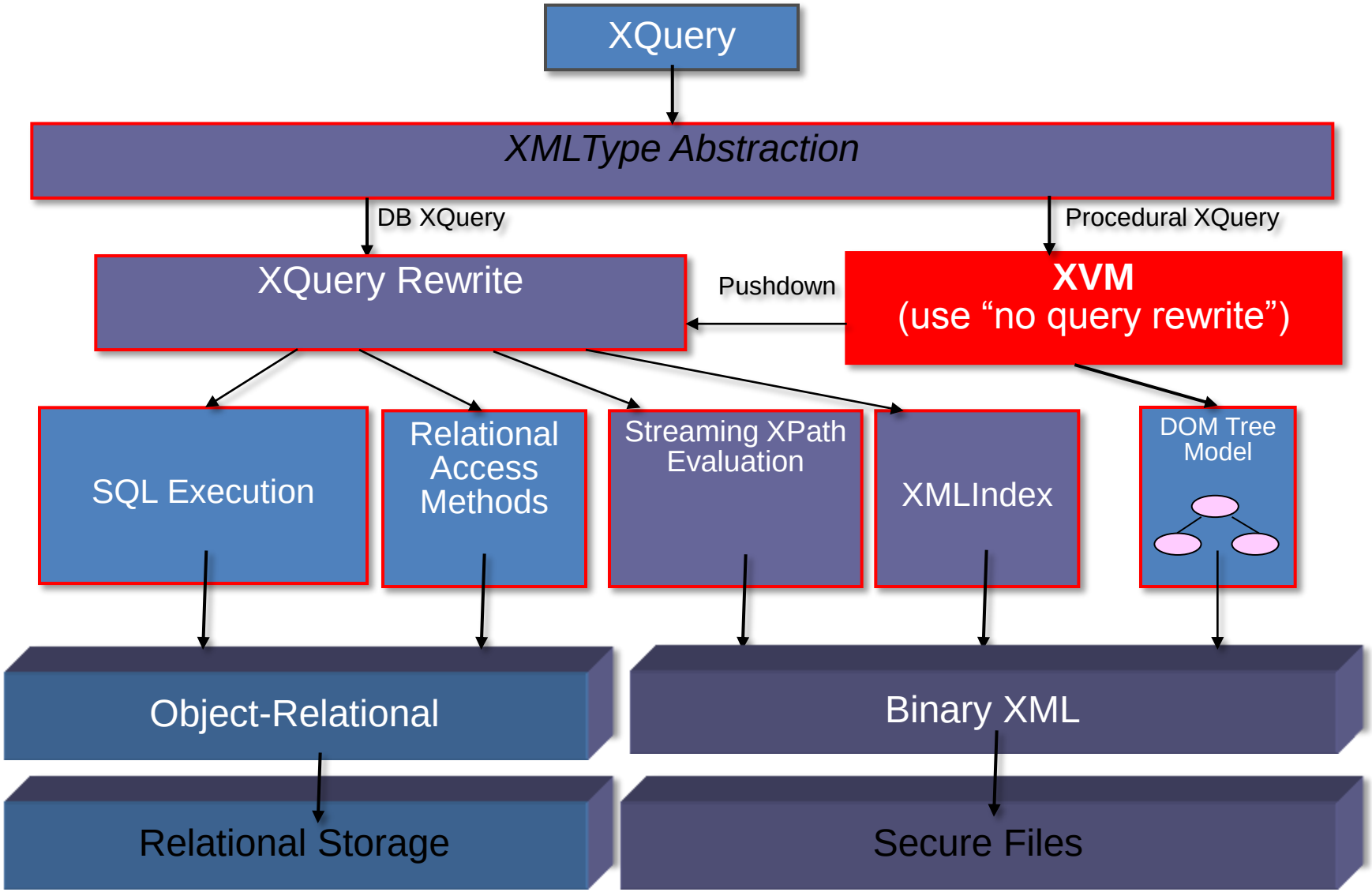
Building (streaming) Bridges



Oracle **XML DB**



- NO cost option
- C *(native / embedded kernel)*
- (XQuery) Standards
- Code maintained by Oracle



Source: S317428: Building Really Scalable XML Applications with Oracle XML DB and Oracle Text

So about what are we talking ?

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Deutsch
Français
Português
Русский

Andrea Andreani

From Wikipedia, the free encyclopedia

Andrea Andreani (1540–1623) was an Italian engraver on wood, who was among the first printmakers in Italy to use chiaroscuro, which required multiple colours.

Born and generally active in Mantua about 1540 (Brulliot says 1560) and died at Rome in 1623. His engravings are scarce and valuable, and are chiefly copies of Mantegna, Albrecht Dürer, Parmigianino and Titian. The most remarkable of his works are *Mercury and Ignorance*, the *Deluge*, *Pharaoh's Host Drowned in the Red Sea* (after Titian), the *Triumph of Caesar* (after Mantegna), and *Christ retiring from the judgment-seat of Pilate* after a relief by Giambologna. He was active 1584-1610 in Florence.[1]

Triumphus Caesari, by Andreani, after a painting by Mantegna

References

[edit]

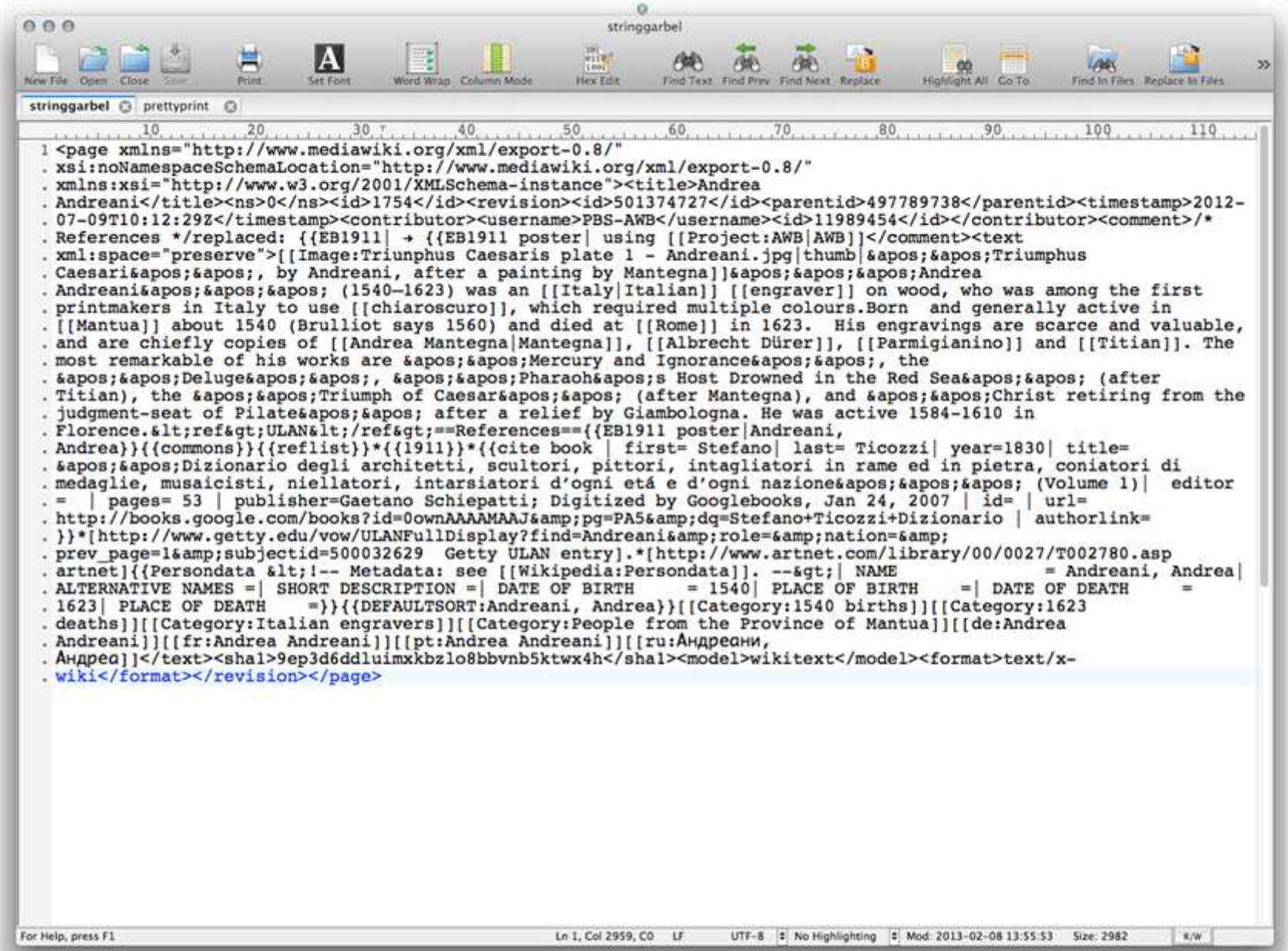
1. ^ ULAN

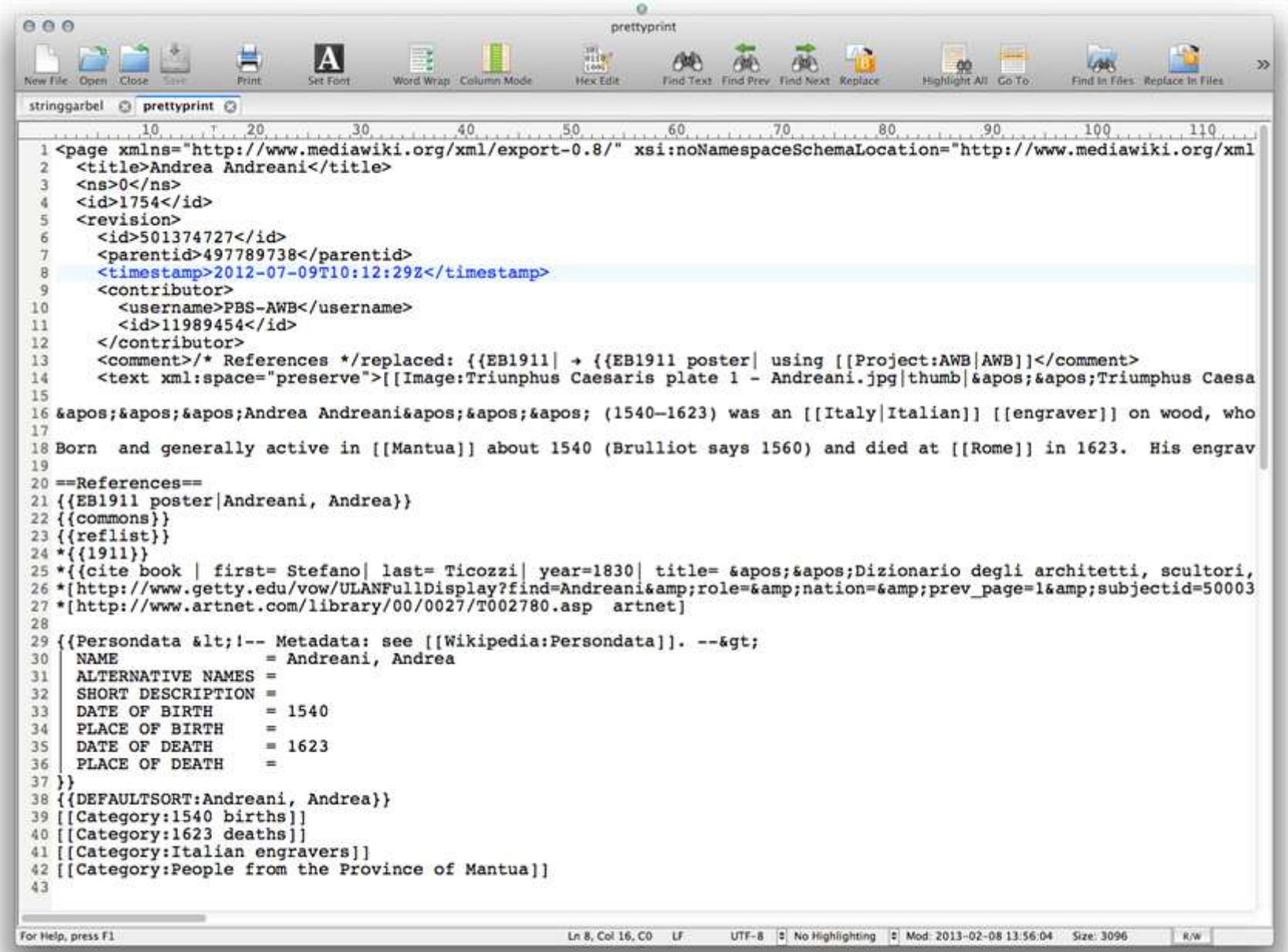
- This article incorporates text from a publication now in the public domain: Chisholm, Hugh, ed. (1911). *Encyclopædia Britannica* (11th ed.). Cambridge University Press.
- Ticozzi, Stefano (1830). Dizionario degli architetti, scultori, pittori, intagliatori in rame ed in pietra, coniatori di medaglie, musaicisti, niellatori, intarsiatori d'ogni etá e d'ogni nazione' *(Volume 1)*. Gaetano Schiepatti; Digitized by Googlebooks, Jan 24, 2007. pp. 53.
- Getty ULAN entry.
- artnet

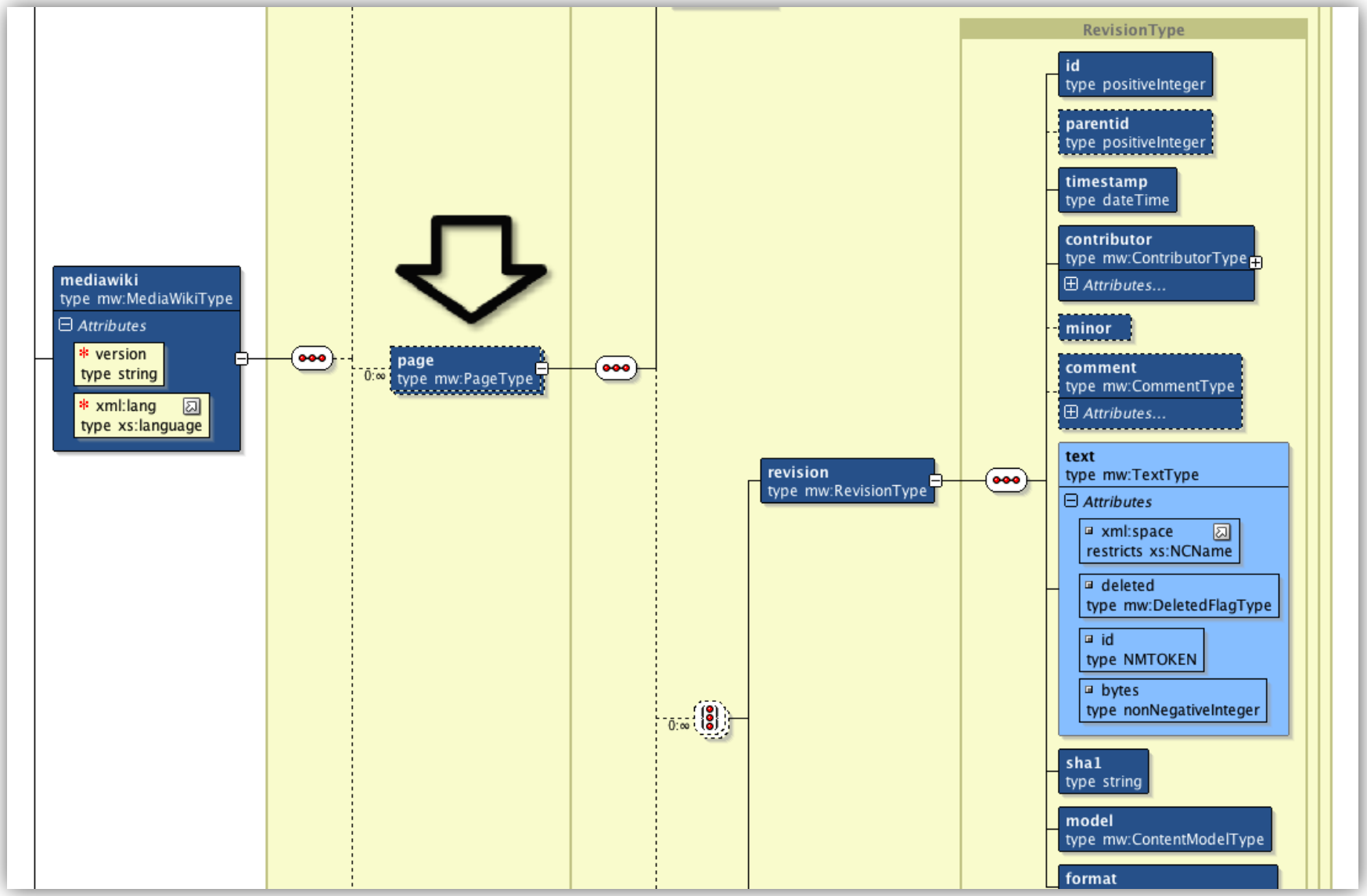
Wikisource has the text of the 1911 Encyclopædia Britannica article ***Andreani, Andrea.***

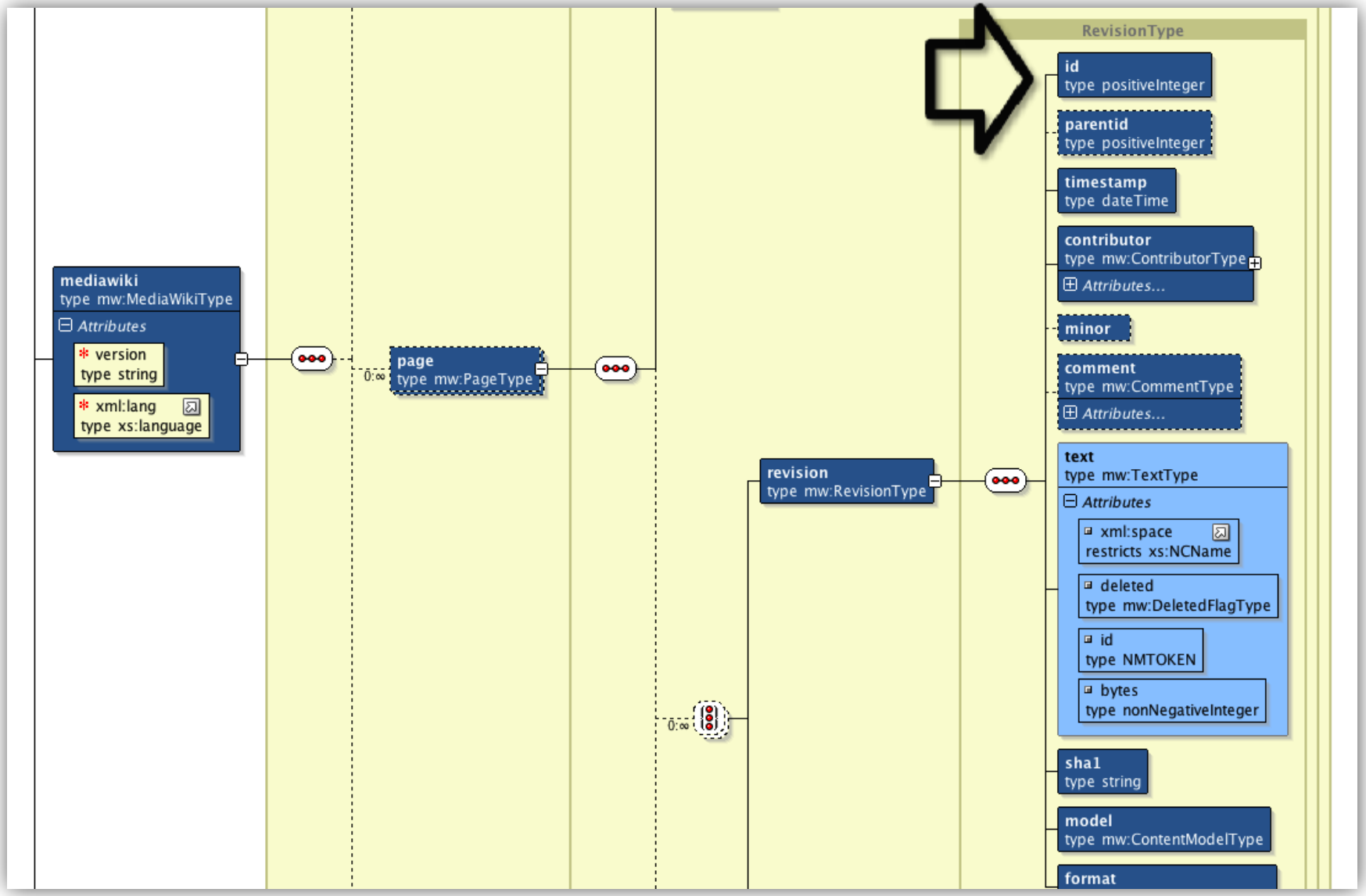
Wikimedia Commons has media related to: ***Andrea Andreani***

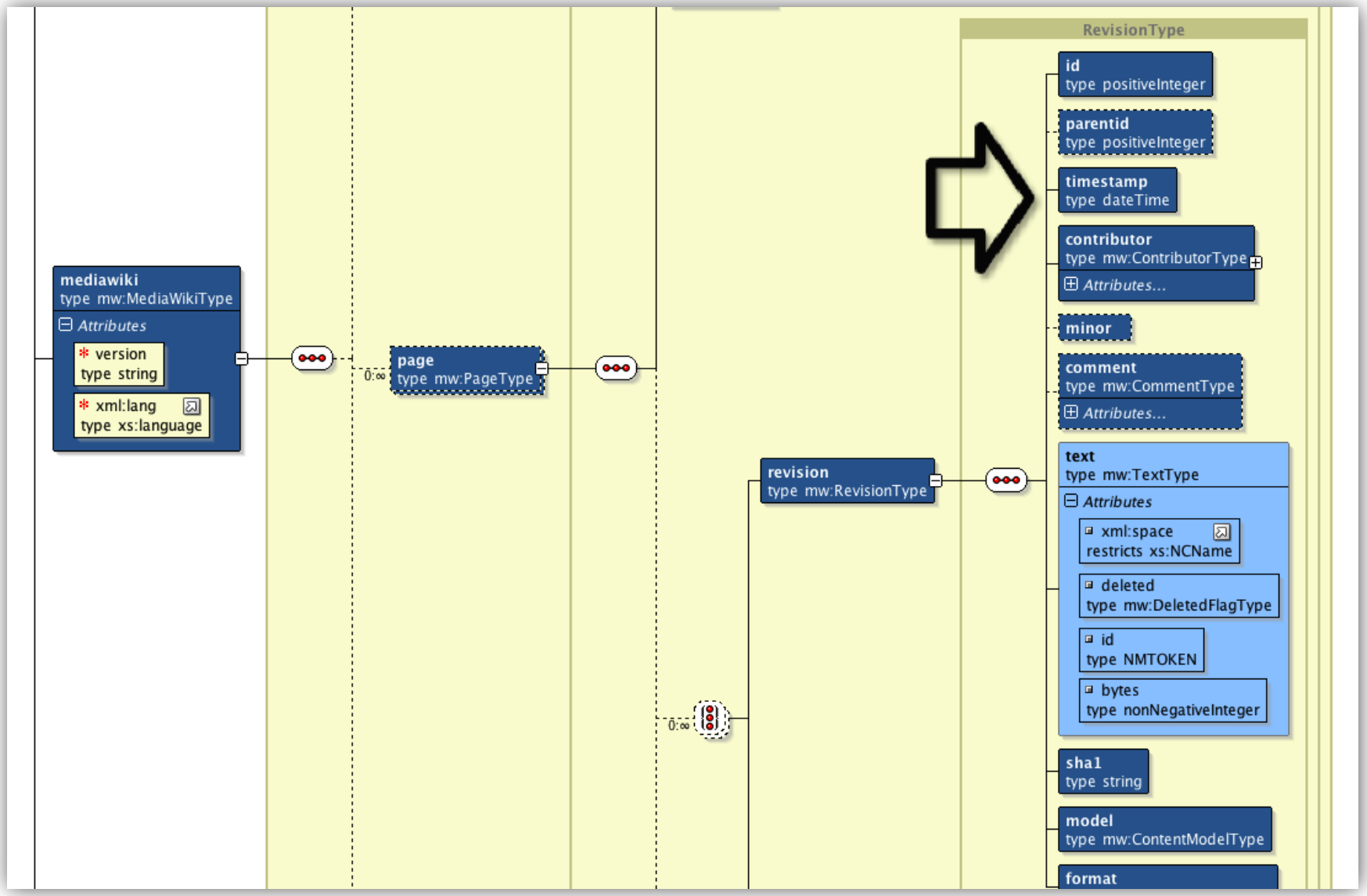
Categories: 1540 births | 1623 deaths | Italian engravers | People from the Province of Mantua

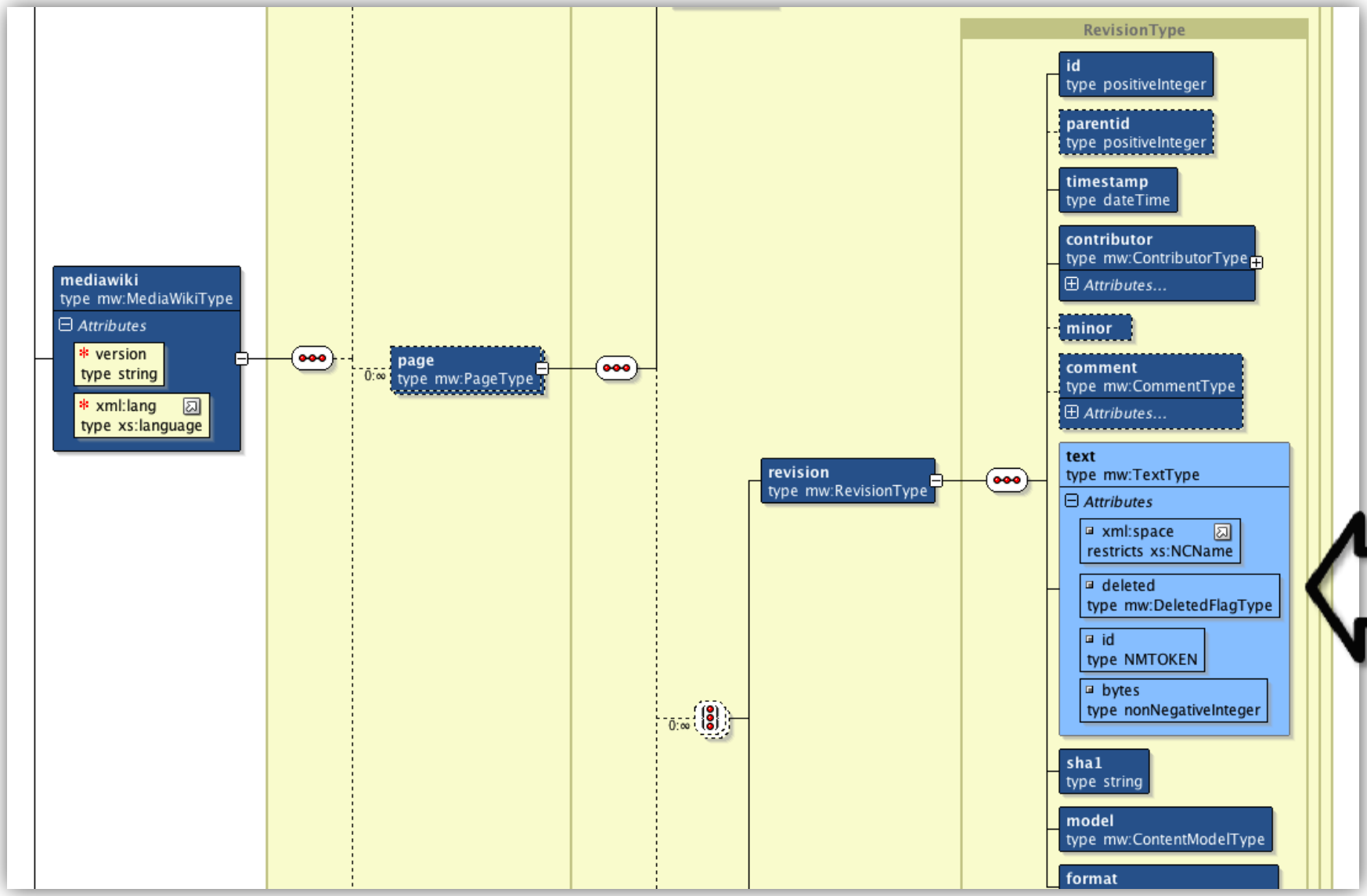

stringgarbel
New File
Open
Close
Save
Print
Set Font
Word Wrap
Column Mode
Hex Edit
Find Text
Find Prev
Find Next
Replace
Highlight All
Go To
Find In Files
Replace In Files
stringgarbel
prettyprint
10
20
30
40
50
60
70
80
90
100
110
1 <page xmlns="http://www.mediawiki.org/xml/export-0.8/"
. xsi:noNamespaceSchemaLocation="http://www.mediawiki.org/xml/export-0.8/"
. xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"><title>Andrea
. Andreani</title><ns>0</ns><id>1754</id><revision><id>501374727</id><parentid>497789738</parentid><timestamp>2012-
. 07-09T10:12:29Z</timestamp><contributor><username>PBS-AWB</username><id>11989454</id></contributor><comment>/*
. References */replaced: {{EB1911| → {{EB1911 poster| using [[Project:AWB|AWB]]</comment><text
. xml:space="preserve">[[Image:Triunphus Caesaris plate 1 - Andreani.jpg|thumb|''Triumphus
. Caesari'', by Andreani, after a painting by Mantegna]]'''Andrea
. Andreani''' (1540–1623) was an [[Italy|Italian]] [[engraver]] on wood, who was among the first
. printmakers in Italy to use [[chiaroscuro]], which required multiple colours.Born and generally active in
. [[Mantua]] about 1540 (Brulliot says 1560) and died at [[Rome]] in 1623. His engravings are scarce and valuable,
. and are chiefly copies of [[Andrea Mantegna|Mantegna]], [[Albrecht Dürer]], [[Parmigianino]] and [[Titian]]. The
. most remarkable of his works are ''Mercury and Ignorance'', the
. ''Deluge'', ''Pharaoh's Host Drowned in the Red Sea'' (after
. Titian), the ''Triumph of Caesar'' (after Mantegna), and ''Christ retiring from the
. judgment-seat of Pilate'' after a relief by Giambologna. He was active 1584-1610 in
. Florence.<ref>ULAN</ref>==References=={{EB1911 poster|Andreani,
. Andrea}}{{commons}}{{reflist}}*{{1911}}*{{cite book | first= Stefano| last= Ticozzi| year=1830| title=
. ''Dizionario degli architetti, scultori, pittori, intagliatori in rame ed in pietra, coniatori di
. medaglie, musaicisti, niellatori, intarsiatori d'ogni età e d'ogni nazione''' (Volume 1)| editor
. = | pages= 53 | publisher=Gaetano Schiepatti; Digitized by Googlebooks, Jan 24, 2007 | id= | url=
. http://books.google.com/books?id=0ownAAAAMAAJ&pg=PA5&dq=Stefano+Ticozzi+Dizionario | authorlink=
. }}*[http://www.getty.edu/vow/ULANFullDisplay?find=Andreani&role=&nation=&
. prev_page=1&subjectid=500032629 Getty ULAN entry].*[http://www.artnet.com/library/00/0027/T002780.asp
. artnet]{{Persondata <!-- Metadata: see [[Wikipedia:Persondata]]. -->| NAME = Andreani, Andrea|
. ALTERNATIVE NAMES =| SHORT DESCRIPTION =| DATE OF BIRTH = 1540| PLACE OF BIRTH =| DATE OF DEATH =
. 1623| PLACE OF DEATH =}}{{DEFAULTSORT:Andreani, Andrea}}[[Category:1540 births]][[Category:1623
. deaths]][[Category:Italian engravers]][[Category:People from the Province of Mantua]][[de:Andrea
. Andreani]][[fr:Andrea Andreani]][[pt:Andrea Andreani]][[ru:Андреани,
. Андреа]]</text><sha1>9ep3d6ddluimxkbzlo8bbvnb5ktwx4h</sha1><model>wikitext</model><format>text/x-
. wiki</format></revision></page>
For Help, press F1
Ln 1, Col 2959, C0
LF
UTF-8
No Highlighting
Mod: 2013-02-08 13:55:53
Size: 2982
R/W


prettyprint
New File
Open
Close
Save
Print
Set Font
Word Wrap
Column Mode
Hex Edit
Find Text
Find Prev
Find Next
Replace
Highlight All
Go To
Find In Files
Replace In Files
stringgarbel
prettyprint
<page xmlns="http://www.mediawiki.org/xml/export-0.8/" xsi:noNamespaceSchemaLocation="http://www.mediawiki.org/xml
<title>Andrea Andreani</title>
<ns>0</ns>
<id>1754</id>
<revision>
<id>501374727</id>
<parentid>497789738</parentid>
<timestamp>2012-07-09T10:12:29Z</timestamp>
<contributor>
<username>PBS-AWB</username>
<id>11989454</id>
</contributor>
<comment>/* References */replaced: {{EB1911| → {{EB1911 poster| using [[Project:AWB|AWB]]</comment>
<text xml:space="preserve">[[Image:Triunphus Caesaris plate 1 - Andreani.jpg|thumb|''Triumphus Caesa
'''Andrea Andreani''' (1540–1623) was an [[Italy|Italian]] [[engraver]] on wood, who
Born and generally active in [[Mantua]] about 1540 (Brulliot says 1560) and died at [[Rome]] in 1623. His engrav
==References==
{{EB1911 poster|Andreani, Andrea}}
{{commons}}
{{reflist}}
*{{1911}}
*{{cite book | first= Stefano| last= Ticozzi| year=1830| title= ''Dizionario degli architetti, scultori,
*[http://www.getty.edu/vow/ULANFullDisplay?find=Andreani&role=&nation=&prev_page=1&subjectid=50003
*[http://www.artnet.com/library/00/0027/T002780.asp artnet]
{{Persondata <!-- Metadata: see [[Wikipedia:Persondata]]. -->
NAME = Andreani, Andrea
ALTERNATIVE NAMES =
SHORT DESCRIPTION =
DATE OF BIRTH = 1540
PLACE OF BIRTH =
DATE OF DEATH = 1623
PLACE OF DEATH =
}}
{{DEFAULTSORT:Andreani, Andrea}}
[[Category:1540 births]]
[[Category:1623 deaths]]
[[Category:Italian engravers]]
[[Category:People from the Province of Mantua]]
For Help, press F1
Ln 8, Col 16, C0
LF
UTF-8
No Highlighting
Mod: 2013-02-08 13:56:04
Size: 3096
R/W


RevisionType
mediawiki
type mw:MediaWikiType
Attributes
* version
type string
* xml:lang
type xs:language
page
type mw:PageType
0:∞
revision
type mw:RevisionType
0:∞
id
type positiveInteger
parentid
type positiveInteger
timestamp
type dateTime
contributor
type mw:ContributorType
Attributes...
minor
comment
type mw:CommentType
Attributes...
text
type mw:TextType
Attributes
xml:space
restricts xs:NCName
deleted
type mw:DeletedFlagType
id
type NMTOKEN
bytes
type nonNegativeInteger
sha1
type string
model
type mw:ContentModelType
format


RevisionType
mediawiki
type mw:MediaWikiType
Attributes
* version
type string
* xml:lang
type xs:language
page
type mw:PageType
0:∞
revision
type mw:RevisionType
0:∞
id
type positiveInteger
parentid
type positiveInteger
timestamp
type dateTime
contributor
type mw:ContributorType
Attributes...
minor
comment
type mw:CommentType
Attributes...
text
type mw:TextType
Attributes
xml:space
restricts xs:NCName
deleted
type mw:DeletedFlagType
id
type NMTOKEN
bytes
type nonNegativeInteger
sha1
type string
model
type mw:ContentModelType
format


RevisionType
mediawiki
type mw:MediaWikiType
Attributes
* version
type string
* xml:lang
type xs:language
page
type mw:PageType
0:∞
revision
type mw:RevisionType
0:∞
id
type positiveInteger
parentid
type positiveInteger
timestamp
type dateTime
contributor
type mw:ContributorType
Attributes...
minor
comment
type mw:CommentType
Attributes...
text
type mw:TextType
Attributes
xml:space
restricts xs:NCName
deleted
type mw:DeletedFlagType
id
type NMTOKEN
bytes
type nonNegativeInteger
sha1
type string
model
type mw:ContentModelType
format


RevisionType
mediawiki
type mw:MediaWikiType
Attributes
* version
type string
* xml:lang
type xs:language
page
type mw:PageType
0:∞
revision
type mw:RevisionType
0:∞
id
type positiveInteger
parentid
type positiveInteger
timestamp
type dateTime
contributor
type mw:ContributorType
Attributes...
minor
comment
type mw:CommentType
Attributes...
text
type mw:TextType
Attributes
xml:space
restricts xs:NCName
deleted
type mw:DeletedFlagType
id
type NMTOKEN
bytes
type nonNegativeInteger
sha1
type string
model
type mw:ContentModelType
format

WikiPedia

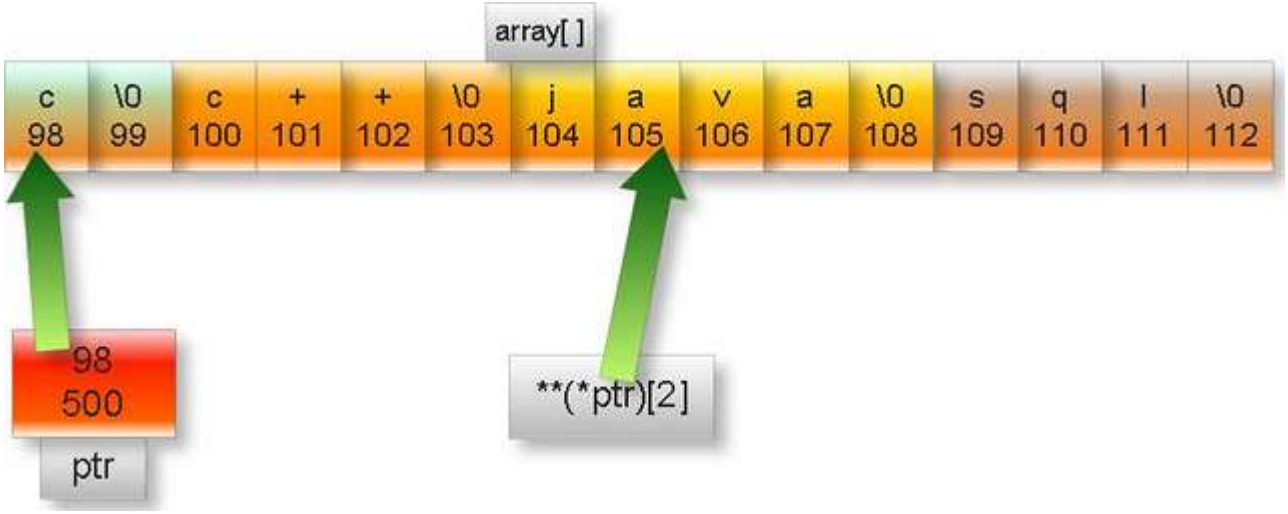
- Structured & Unstructured bits and pieces
- A lot of “unbounded” elements
- Not a lot of restrictions
- The bit with value is in element “tekst”

How do we get this Structured?


CHINA SHIPPING LINE
CSCL SEATTLE

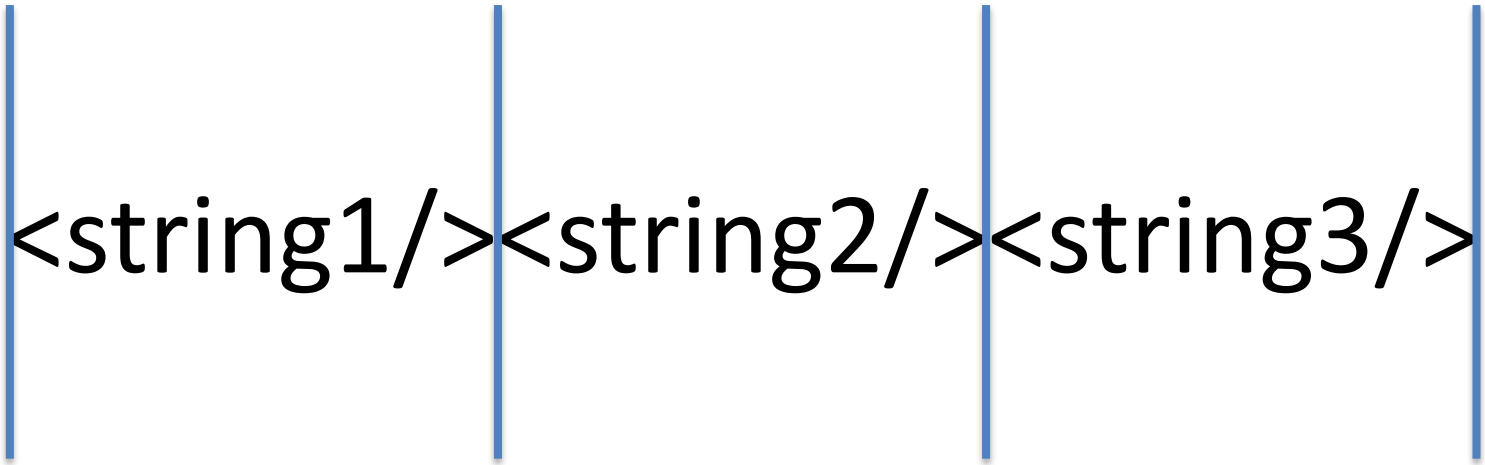

OOCL
OOCL
OOCL
OOCL


50



Strings = small & defined (12c?)

Ename → **pointer += 100;**

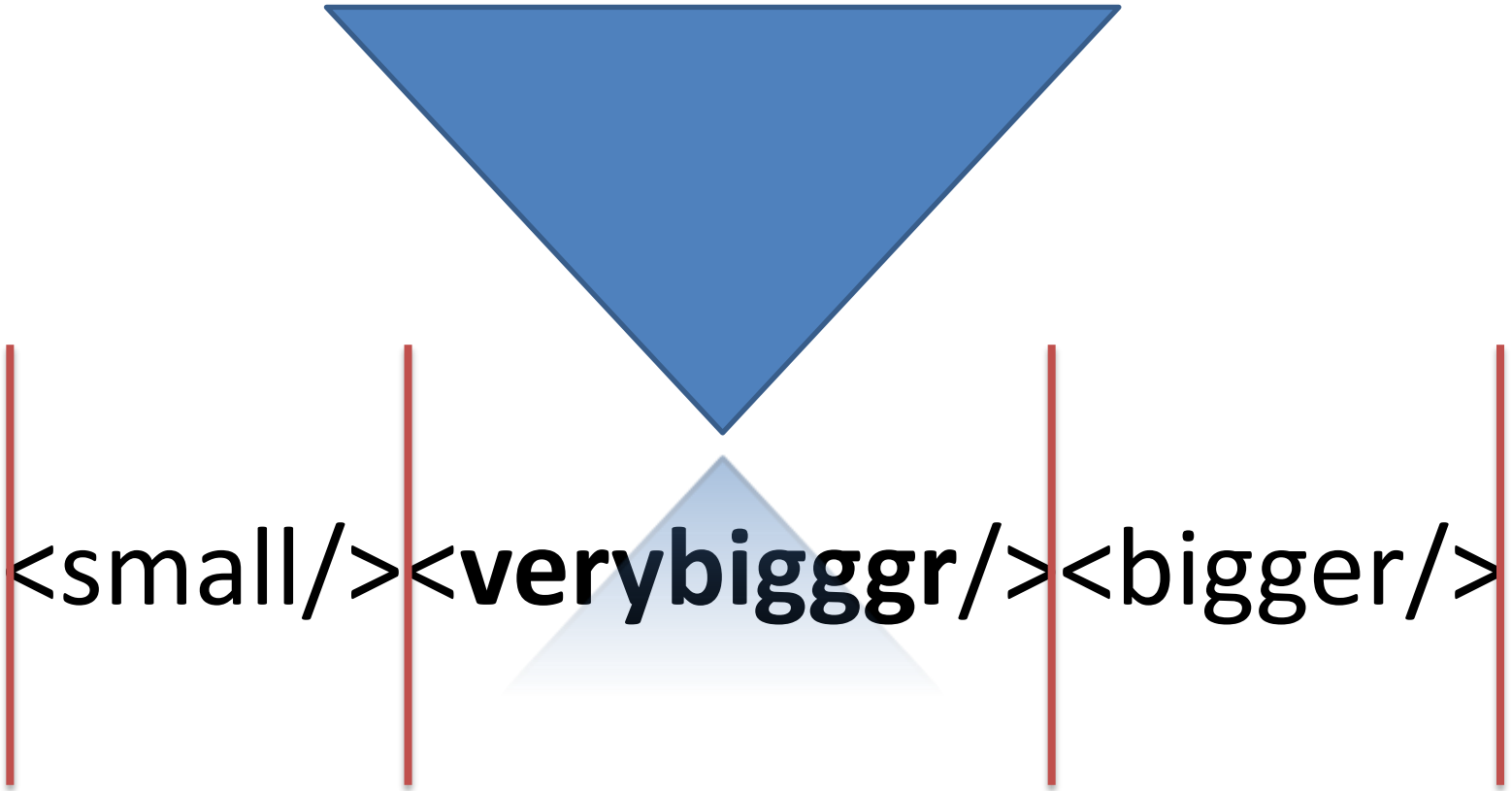

<string1/><string2/><string3/>

Flexible, Humans

No Design Patterns


<small/><verybigggr/><bigger/>

```
<verybigggr>
        <empno>1</empno><ename>Marco</ename>
        <empno>2</empno>
</verybigggr>
```




MAERSK


CHINA SHIPPING LINE
CSCL SEATTLE


OOCL
OOCL
OOCL
OOCL


50

DIVIDE & CONQUER

PLAY

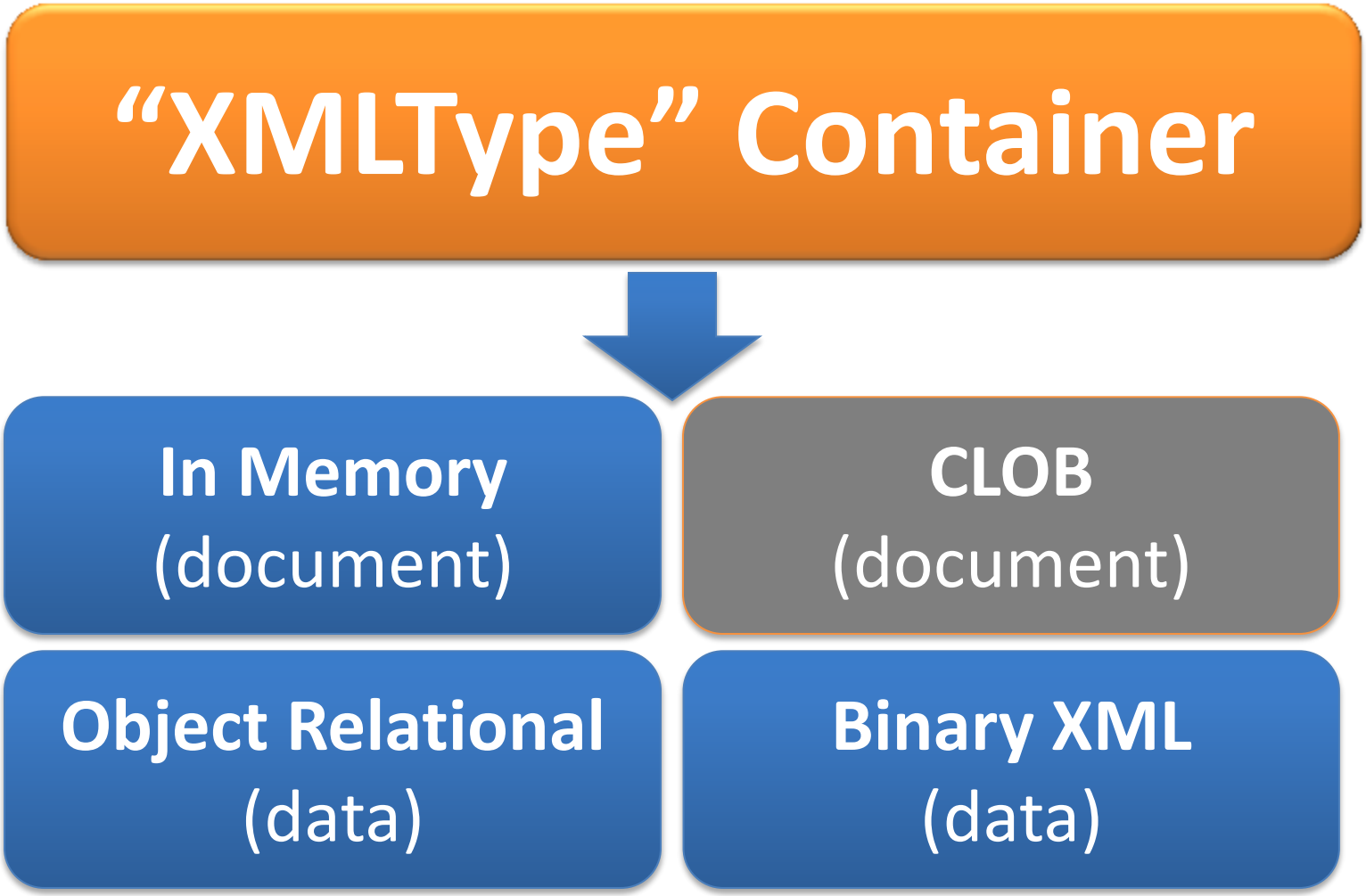
HIGH SCORES

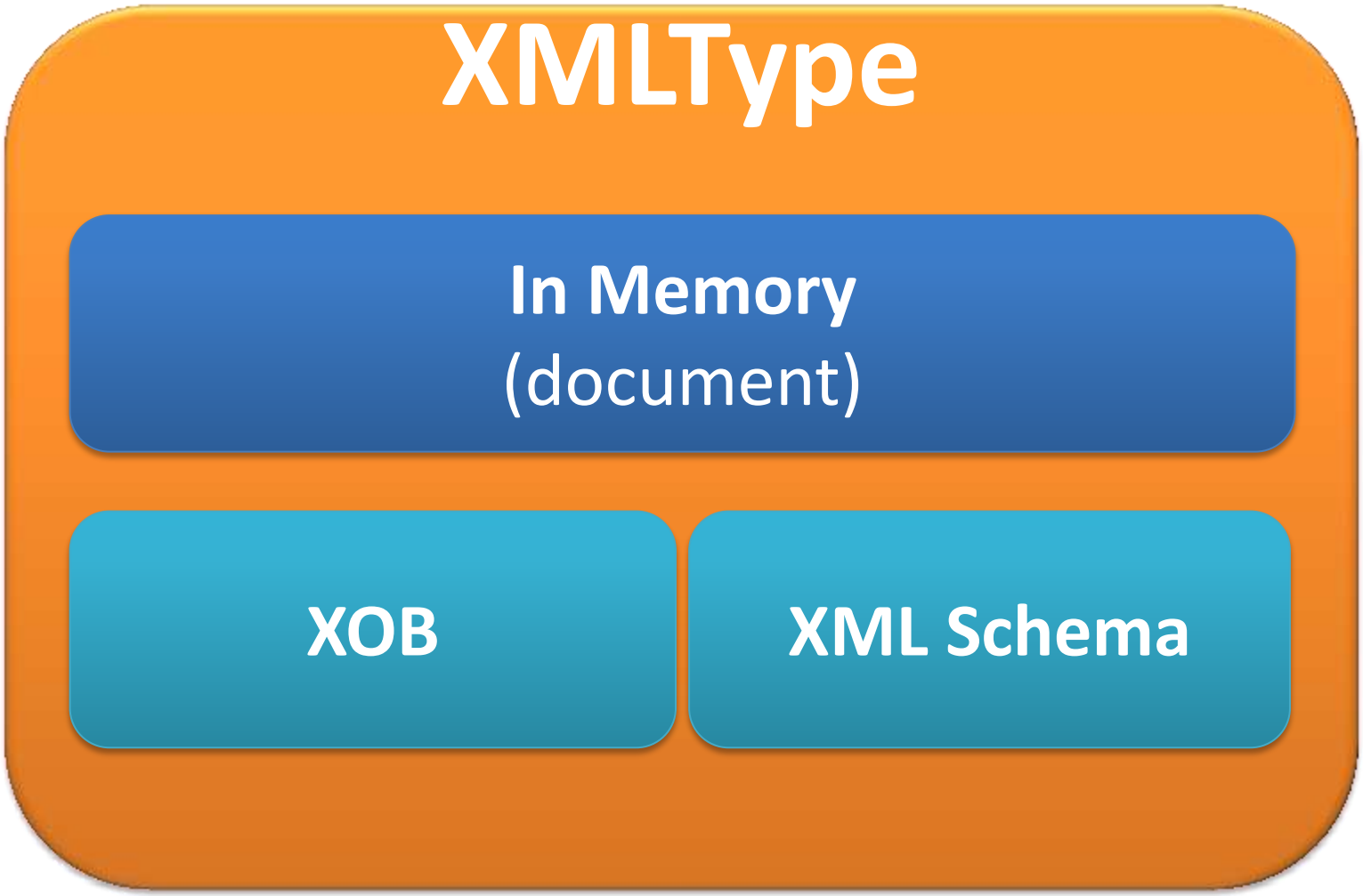
HOW TO PLAY

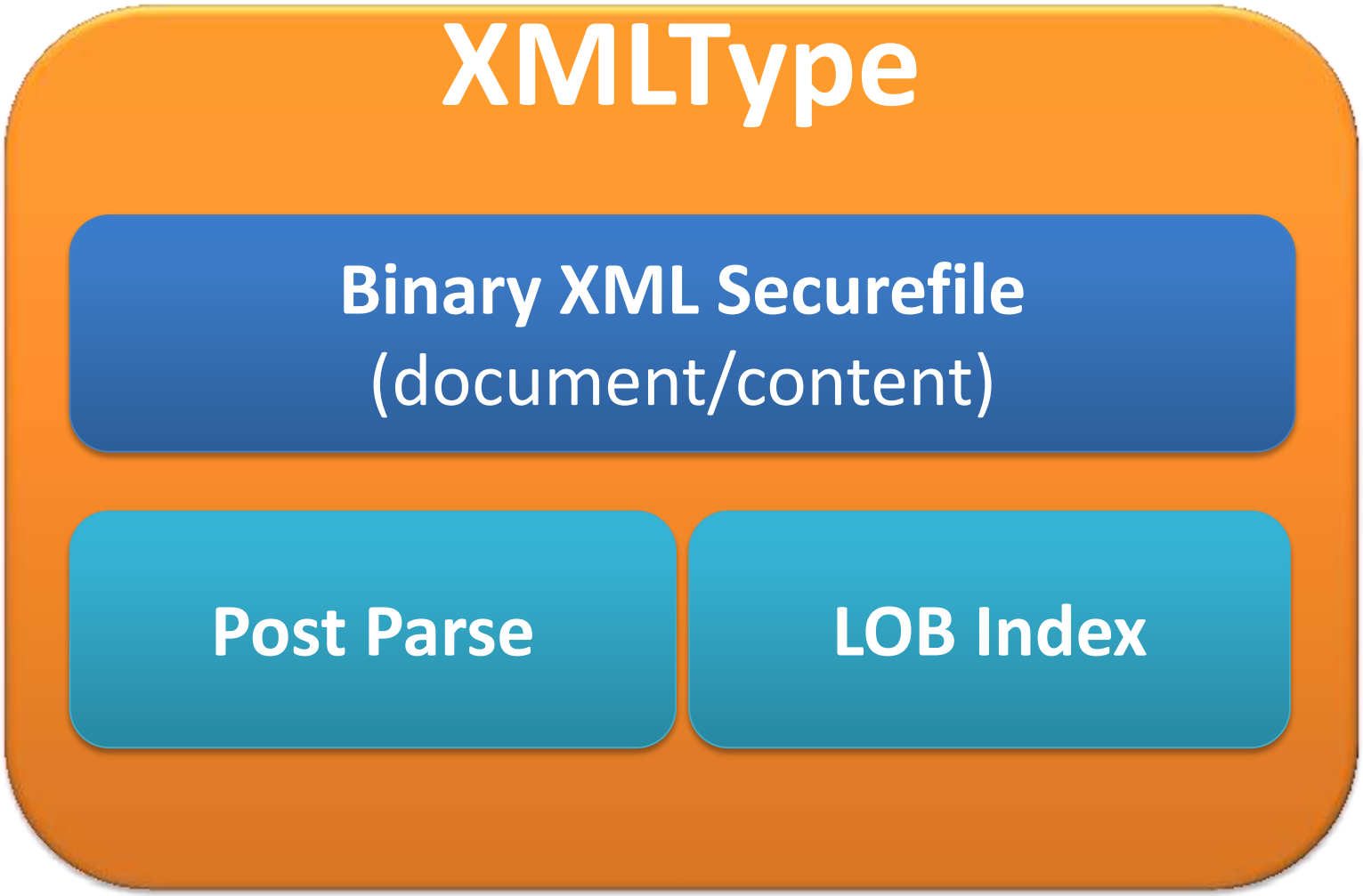
CREDITS

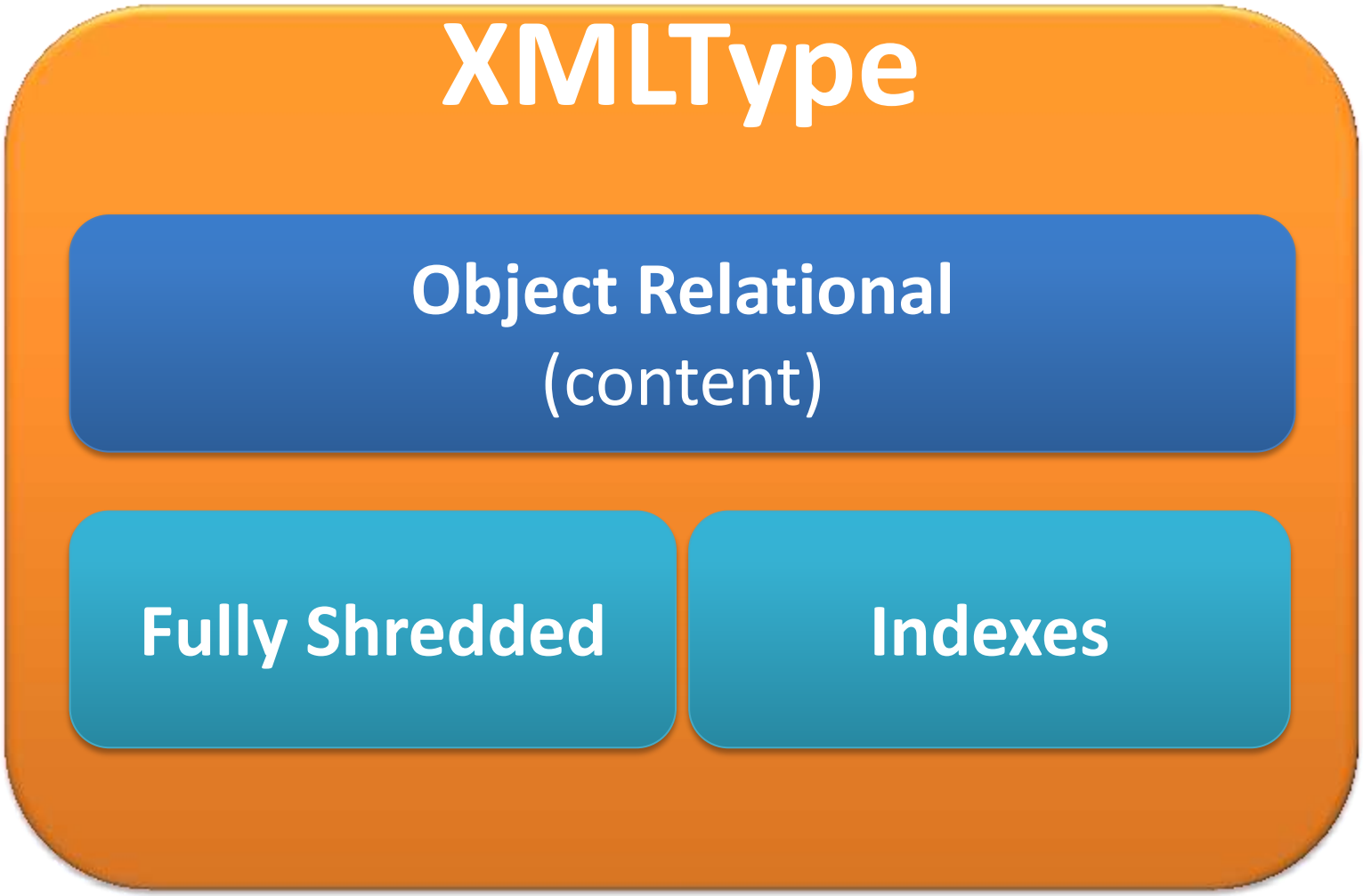
RATE THIS APP

We need options!


"XMLType" Container
In Memory
(document)
CLOB
(document)
Object Relational
(data)
Binary XML
(data)


XMLType
In Memory
(document)
XOB
XML Schema


XMLType
Binary XML Securefile
(document/content)
Post Parse
LOB Index


XMLType
Object Relational
(content)
Fully Shredded
Indexes

Something else to Realize !

“What is the *fastest* way to get this stuff in the database…?”

“…it depends…”

“So what **is** the *fastest* way to get XML in the database…?”

“…it depends…”

"So what **is** the ***fastest*** way to get XML in the database…

… **and** ***useful*** in my case…?"

Garbage IN – Garbage OUT

WikiPedia

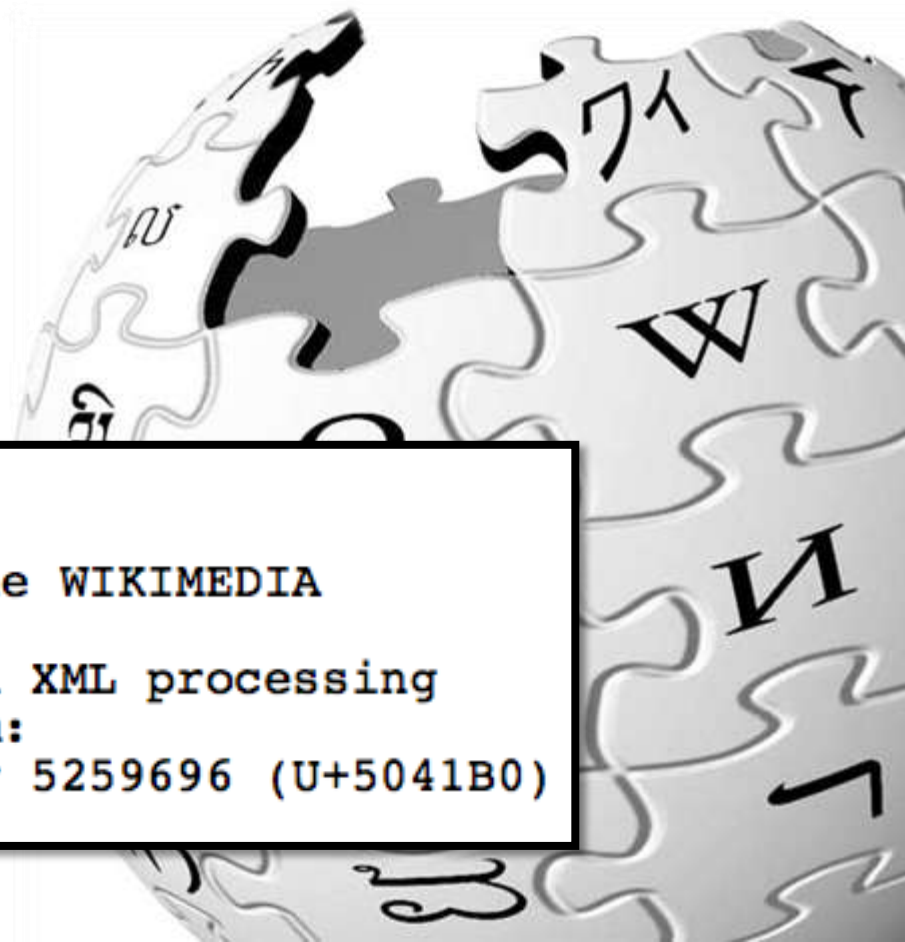
- SQL*Loader
- Parallel or Direct
- Securefile LOB Column
- 2.5 hours

And no (performant) way to get the details out…

a.k.a “completely useless”

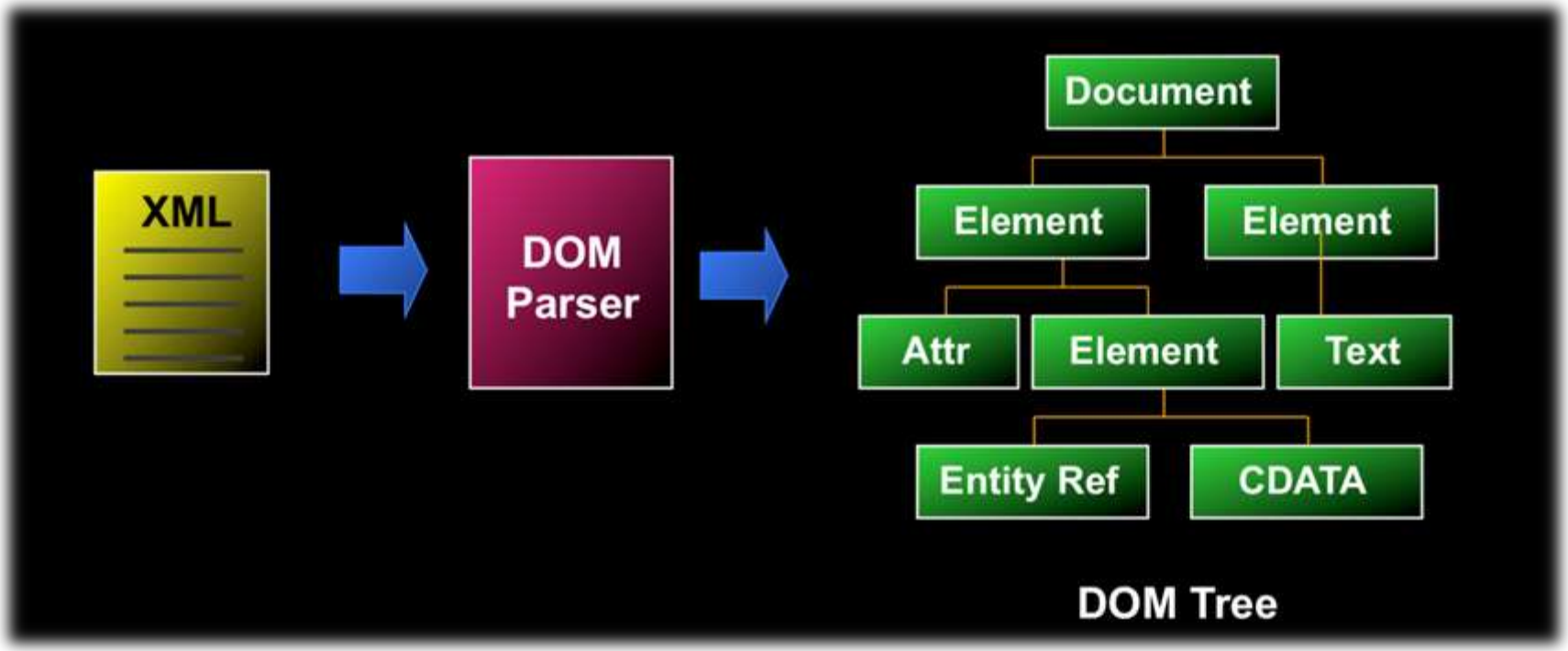
WikiPedia

- SQL*Loader
- Parallel or Direct
- Securefile Binary XML
- …2.5 hours ???

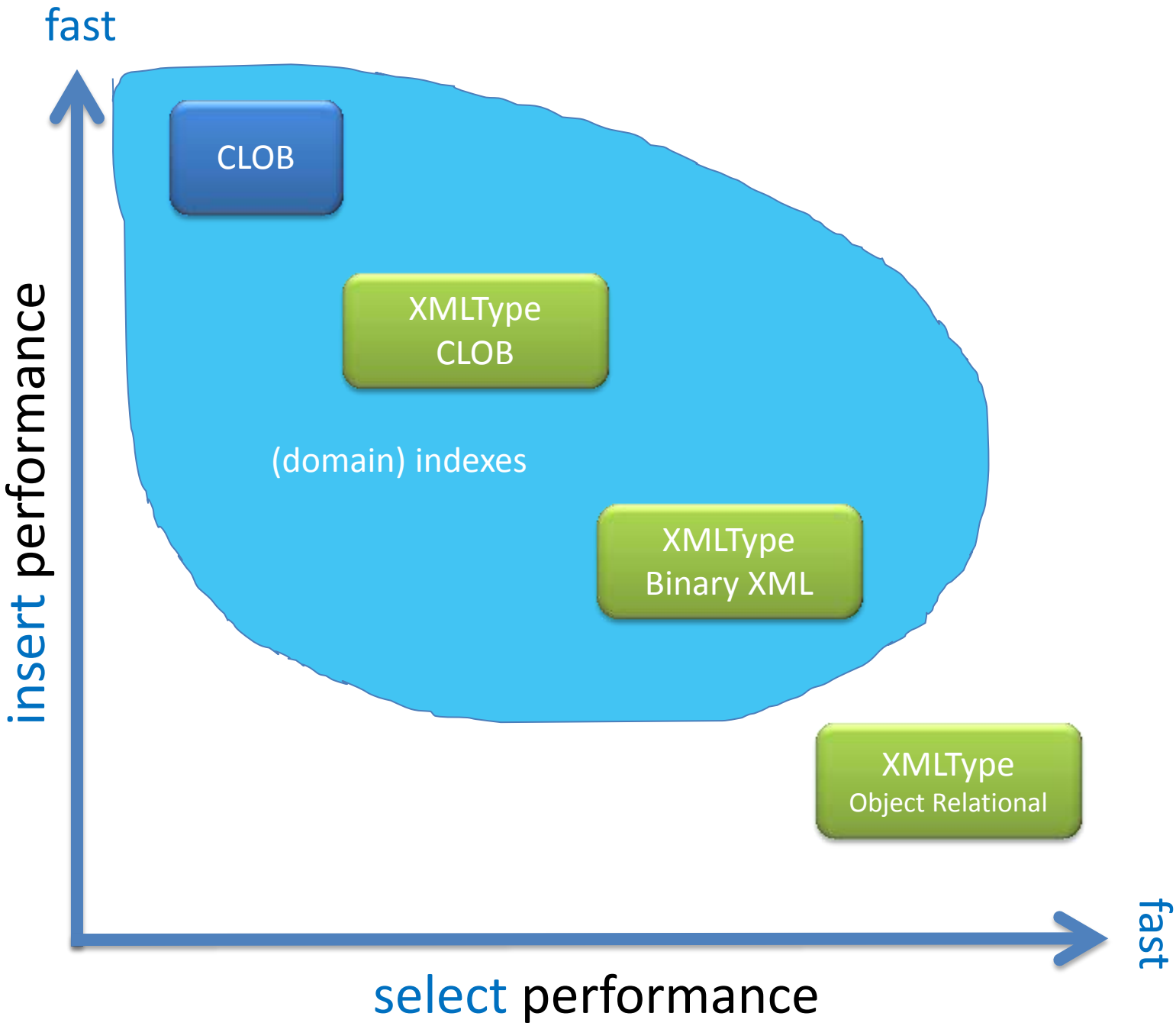


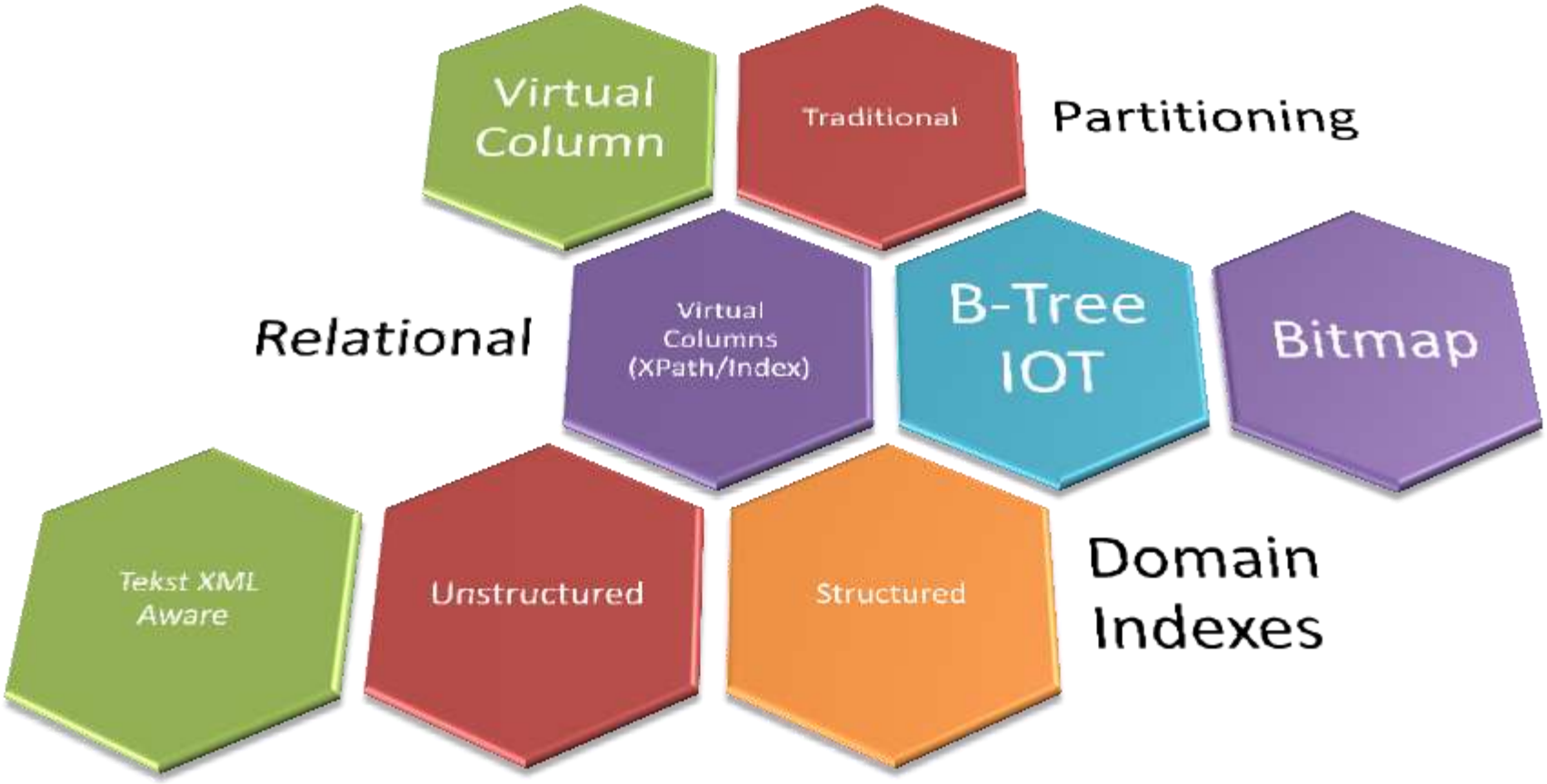
```
XML Parser/LPX error

Parse Error on row 0 in table WIKIMEDIA
OCI-31061: XML event error
OCI-19202: Error occurred in XML processing
In line 1171356 of orastream:
LPX-00217: invalid character 5259696 (U+5041B0)
```

XML Parsing



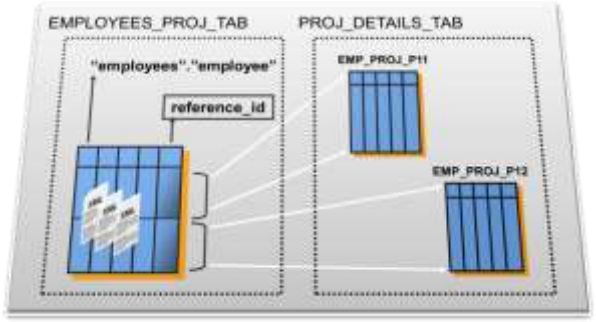
- SAX - Simple API for XML
- DOM - Document Object Module

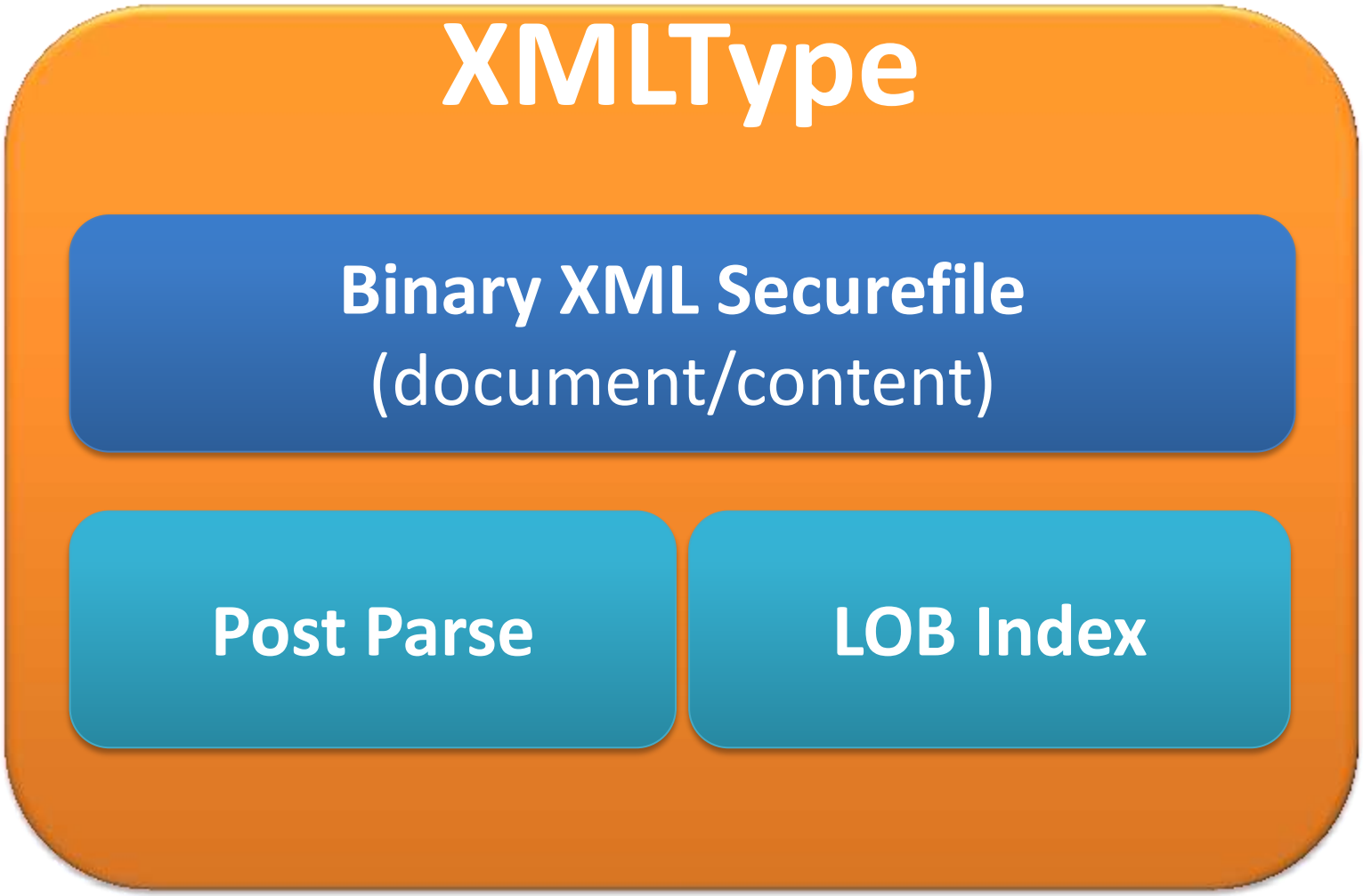

fast
insert performance
CLOB
XMLType
CLOB
(domain) indexes
XMLType
Binary XML
XMLType
Object Relational
fast
select performance


Virtual Column
Traditional
Partitioning
Relational
Virtual Columns (XPath/Index)
B-Tree IOT
Bitmap
Tekst XML Aware
Unstructured
Structured
Domain Indexes

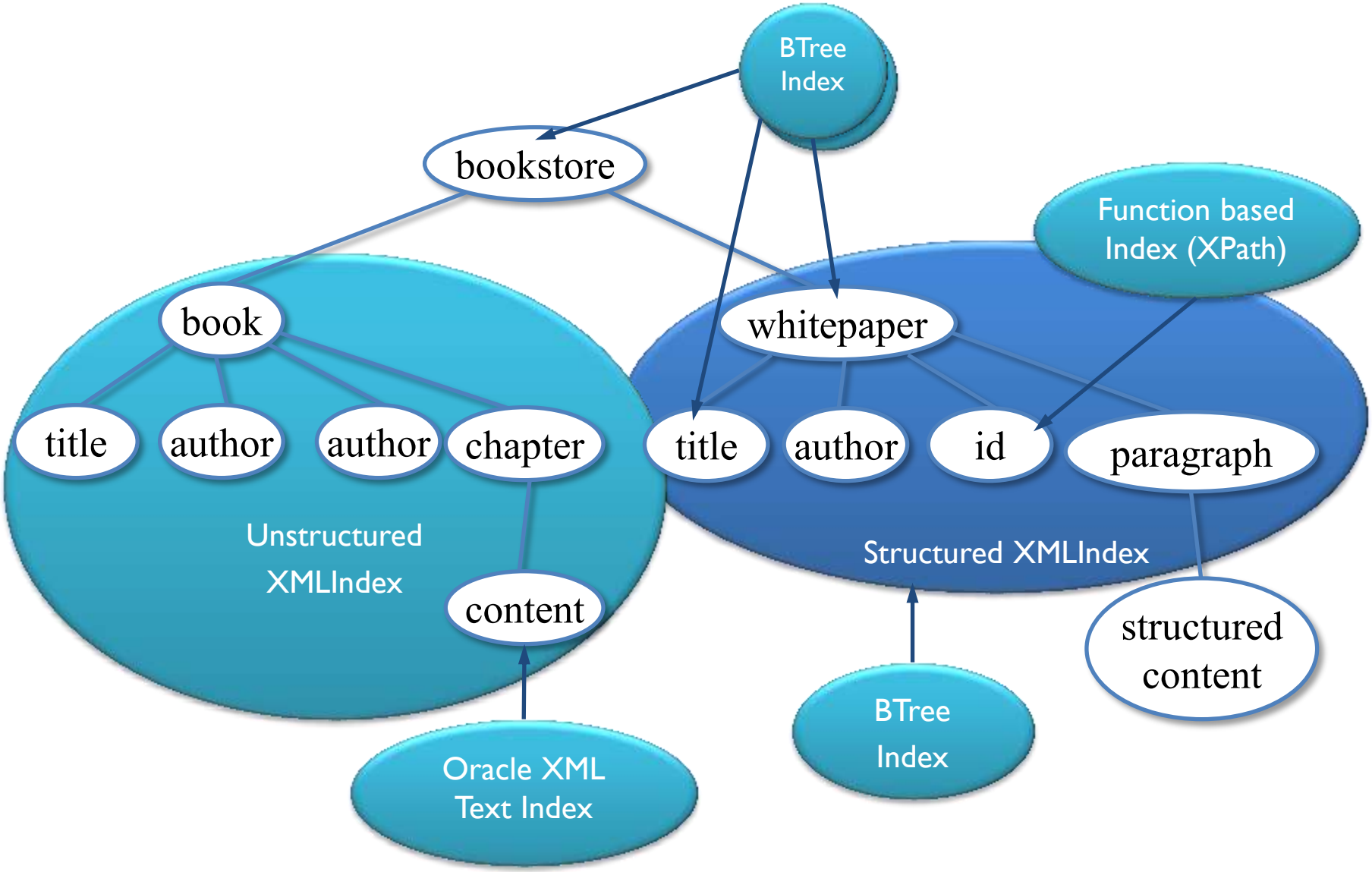
XML Partitioning

- Object Relational Partitioning
  - **Equi-Partitioning** since version Oracle 11.1.0.7.0
- Binary XML Partitioning
  - **Range, List, Hash**
- Local partitioned XMLIndex
  - **LOCAL** keyword in XMLIndex create syntax
- Partition Key on virtual column (Binary XML)
- Partition Key on column (Object Relational)




XMLType
Binary XML Securefile
(document/content)
Post Parse
LOB Index

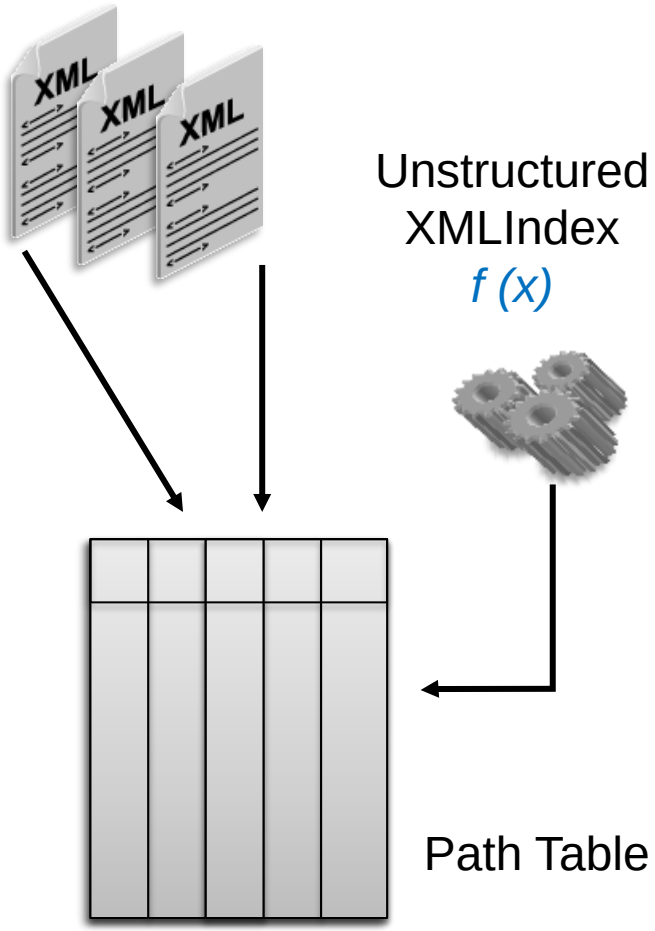
Driving access on CONTENT



Structured Data

Unstructured XMLIndex (UXI)

- **PATH TABLE**
- Use **Path Subsetting**
  - Full Blown XMLIndex can be BIG
- Token Tables (XDB.X$.....)
  - Query re-write on Tokens
  - Fuzzy Searches, //
  - Optimizer Statistics
- Can be maintained **manually**
  - Recorded in Pending Table
- **Secondary** indexes possible



Describe PATH TABLE

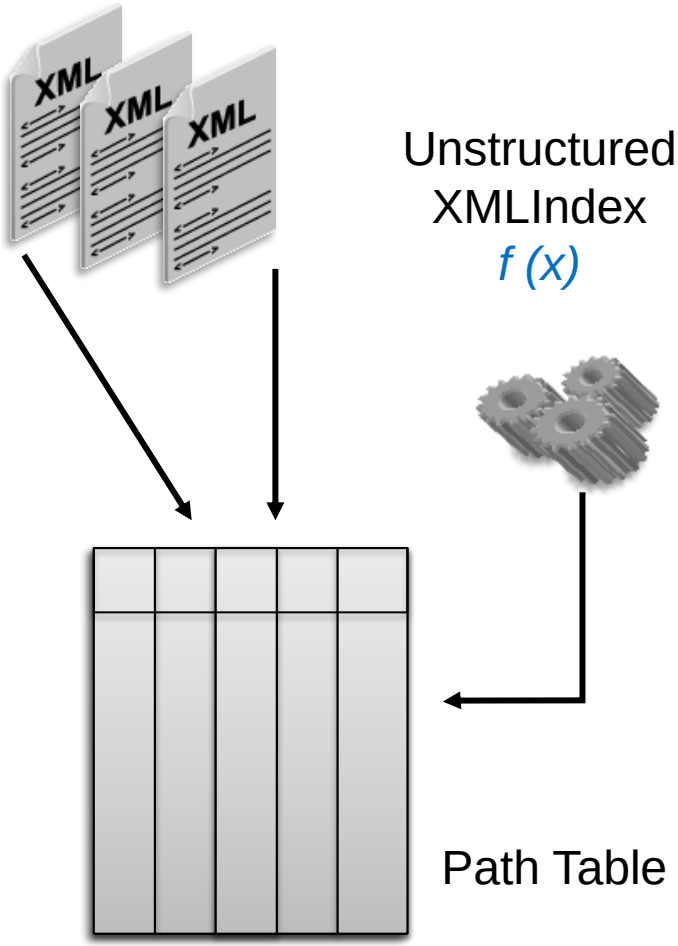
```
SQL> describe UXI_RANGE_PATH_TABLE
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 RID                                                ROWID
 PATHID                                             RAW(8)
 ORDER_KEY                                          RAW(1000)
 LOCATOR                                            RAW(2000)
 VALUE                                              VARCHAR2(4000)
```

```
SQL> select VALUE from UXI_RANGE_PATH_TABLE;
select VALUE from UXI_RANGE_PATH_TABLE
                  *
ERROR at line 1:
ORA-30967: operation directly on the Path Table is disallowed
```

What's hidden…

```
SQL> SELECT * from UXI_RANGE_PATH_TABLE where ROWNUM <= 10;

RID                PATHID ORDER_KEY  LOCATOR                          VALUE
------------------ ------ ---------- -------------------------------- ------------------------------
AAAcB3AABAAARJdAAC 2FD0   0202       010800180402001D00000000000076DF http://www.mediawiki.org/xml/e
                                                                       xport-0.8/

AAAcB3AABAAARJdAAC 47B4   0204       0108001804020020000000000000076DF http://www.w3.org/2001/XMLSche
                                                                       ma-instance

AAAcB3AABAAARJdAAC 7EE9   0206       0108001804020023000000000004120 Ronny and the Daytonas
AAAcB3AABAAARJdAAC 3BB7   0208       010800180402003D0000000000000A50 0
AAAcB3AABAAARJdAAC 0936   020A       0108001804020042000000000008C5 167379
AAAcB3AABAAARJdAAC 15FC   020C02     010800180402004F0000000000002D6 Ronny & the Daytonas
AAAcB3AABAAARJdAAC 3ABC   020C       011000180402004C000000000007861
AAAcB3AABAAARJdAAC 2668   020E02     010800180402006B0000000000008C5 16054053
AAAcB3AABAAARJdAAC 7878   020E04     0108001804020077000000000000673 2003-01-09T23:56:47Z
AAAcB3AABAAARJdAAC 41A2   020E0602   0108001804020092000000000038D3 TUF-KAT

10 rows selected.
```

PATH TABLE



INDEXED COLUMNS

PATH INDEX

- (**PATHID, RID**), BTREE

ORDER INDEX

- (**RID, ORDER_KEY**), BTREE

VALUE INDEX

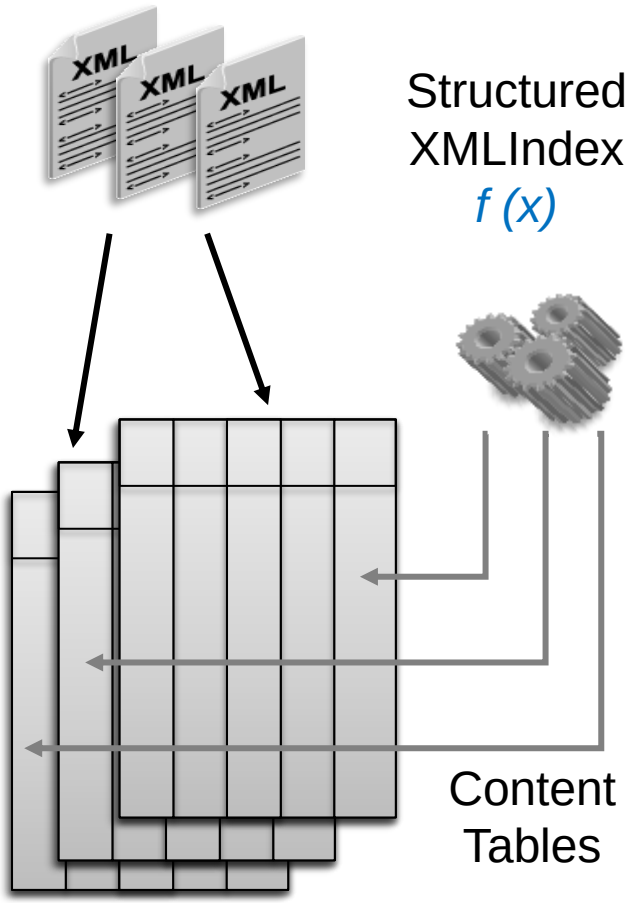
- (SUBSTRB("**VALUE**",1,1599))
- FUNCTION BASED

Not Indexed: LOCATOR column, pointer to XML fragments (XDB.X$...)

SECONDARY INDEXES

Structured XMLIndex (SXI)

- **CONTENT TABLE**(s)
- Based on **XMLTABLE** syntax
- XMLTable construct can be nested:
  - VIRTUAL column alias
- Can be maintained **manually**
- **Secondary** indexes possible

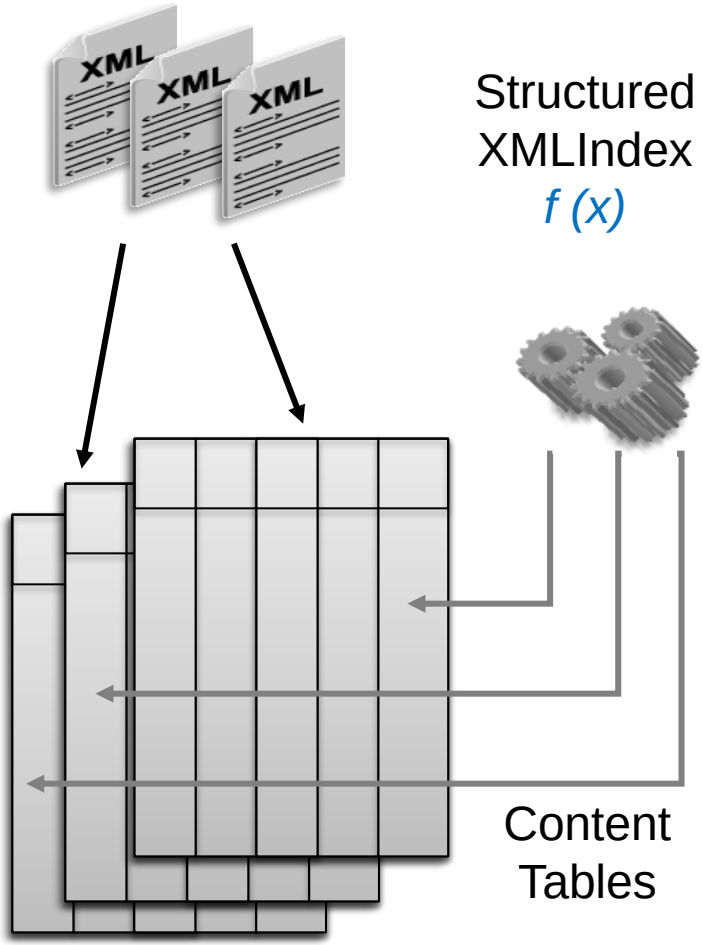


Describe CONTENT TABLE

```
SQL> describe WIKI_SXI_TABLE
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 KEY                                                RAW(1000)
 RID                                                ROWID
 PAGE_ID                                            NUMBER(38)
 PAGE_TITLE                                         VARCHAR2(4000)
 PAGE_REV_TIMESTAMP                                 TIMESTAMP(6) WITH TIME ZONE
```

- A “regular” heap table with columns…
- Ideal for secondary indexes, if needed.

CONTENT TABLE(s)



INDEXED COLUMNS

KEY INDEX

- **(KEY)**, Unique BTREE

RID INDEX

- **(RID)**, Non-Unique BTREE

Indexes needed for combined XMLIndex Types

Mixing Unstructured and Structured XMLIndexes

Your defined columns

Secondary indexes

Binary XML – No Index

Worksheet | Query Builder

```
1  SELECT PAGE_ID, -- NO XMLINDEX --
2         PAGE_TITLE,
3         PAGE_REV_TIMESTAMP
4  FROM   BINARYXML_SECUREFILE_XSD t1,
5         XMLTABLE (xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/' )
6                   , '/page'
7                   PASSING t1.content
8                   COLUMNS
9                     PAGE_ID            NUMBER(9)                    PATH 'id'
10                  , PAGE_TITLE         VARCHAR2(100)                PATH 'title'
11                  , PAGE_REV_TIMESTAMP TIMESTAMP(6) WITH TIME ZONE PATH 'revision/timestamp'
12                 )
13 WHERE PAGE_TITLE='Andrea Andreani'
14 ;
15
16
```

Query Result | Explain Plan

SQL | 0.016 seconds

OPERATION	OBJECT_NAME	OPTIONS	COST
SELECT STATEMENT			39146046
NESTED LOOPS			39146046
TABLE ACCESS	BINARYXML_SECUREFILE_XSD	FULL	25965
XPATH EVALUATION			
Filter Predicates			
CAST(SYS_XQ_UPKXML2SQL(SYS_XQEXVAL(SYS_XQEXT			

Binary XML + XMLIndex (SXI)

Worksheet | Query Builder

```
1  SELECT PAGE_ID, -- WITH XMLINDEX --
2         PAGE_TITLE,
3         PAGE_REV_TIMESTAMP
4  FROM   BINARYXML_TABLE_SECUREFILE t1,
5         XMLTABLE (xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/' )
6                   , '/page'
7                   PASSING t1.content
8                   COLUMNS
9                     PAGE_ID             NUMBER(9)                    PATH 'id'
10                  , PAGE_TITLE          VARCHAR2(100)                PATH 'title'
11                  , PAGE_REV_TIMESTAMP  TIMESTAMP(6) WITH TIME ZONE  PATH 'revision/timestamp'
12                 )
13 WHERE PAGE_TITLE='Andrea Andreani'
14 ;
15
16
```

Query Result | Explain Plan

SQL | 0.014 seconds

OPERATION	OBJECT_NAME	OPTIONS	COST
SELECT STATEMENT			470
PX COORDINATOR			
PX SEND	:TQ10000	QC (RANDOM)	470
NESTED LOOPS			470
PX BLOCK		ITERATOR	
TABLE ACCESS	WIKI_SXI_TABLE	FULL	452
Filter Predicates			
CAST(SYS_SXI_1.PAGE_TITLE AS VARCHAR2			
TABLE ACCESS	BINARYXML_TABLE_SECUREFILE	BY USER ROWID	1

Binary XML + XMLIndex + Sec.Ind.

Worksheet | Query Builder

```
1  SELECT PAGE_ID, -- WITH XMLINDEX and SECONDARY index on CONTENT TABLE COLUMN(s)
2         PAGE_TITLE,
3         PAGE_REV_TIMESTAMP
4  FROM   BINARYXML_TABLE_SECUREFILE t1,
5         XMLTABLE (xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/' )
6                   , '/page'
7                   PASSING t1.content
8                   COLUMNS
9                     PAGE_ID            NUMBER(9)                    PATH 'id'
10                  , PAGE_TITLE         VARCHAR2(100)                PATH 'title'
11                  , PAGE_REV_TIMESTAMP TIMESTAMP(6) WITH TIME ZONE PATH 'revision/timestamp'
12                 )
13 WHERE PAGE_ID=1754
14   AND PAGE_TITLE='Andrea Andreani'
15 ;
16
```

Query Result | Script Output | Query Result 1 | Explain Plan

SQL | 0 seconds

OPERATION	OBJECT_NAME	OPTIONS	COST
SELECT STATEMENT			4
NESTED LOOPS			4
TABLE ACCESS	WIKI_SXI_TABLE	BY INDEX ROWID	3
Filter Predicates			
CAST(SYS_SXI_1.PAGE_TITLE AS VARCHAR2(100)			
INDEX	CNT_PAGE_ID_UXI	UNIQUE SCAN	2
Access Predicates			
SYS_SXI_1.PAGE_ID=1754			
TABLE ACCESS	BINARYXML_TABLE_SECUREFILE	BY USER ROWID	1

Binary XML + XMLIndex + Sec.Ind.

Worksheet | Query Builder

```
1  SELECT PAGE_ID, -- WITH XMLINDEX and SECONDARY index on CONTENT TABLE COLUMN(s)
2         PAGE_TITLE,
3         PAGE_REV_TIMESTAMP
4  FROM   BINARYXML_TABLE_SECUREFILE t1,
5         XMLTABLE (xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/' )
6                   , '/page'
7                   PASSING t1.content
8                   COLUMNS
9                     PAGE_ID            NUMBER(9)                    PATH 'id'
10                  , PAGE_TITLE         VARCHAR2(100)                PATH 'title'
11                  , PAGE_REV_TIMESTAMP TIMESTAMP(6) WITH TIME ZONE PATH 'revision/timestamp'
12                 )
13 WHERE PAGE_ID=1754
14   AND PAGE_TITLE='Andrea Andreani'
15 ;
16
```

Query Result | Explain Plan | Script Output | Query Result 1

SQL | All Rows Fetched: 1 in 0.002 seconds

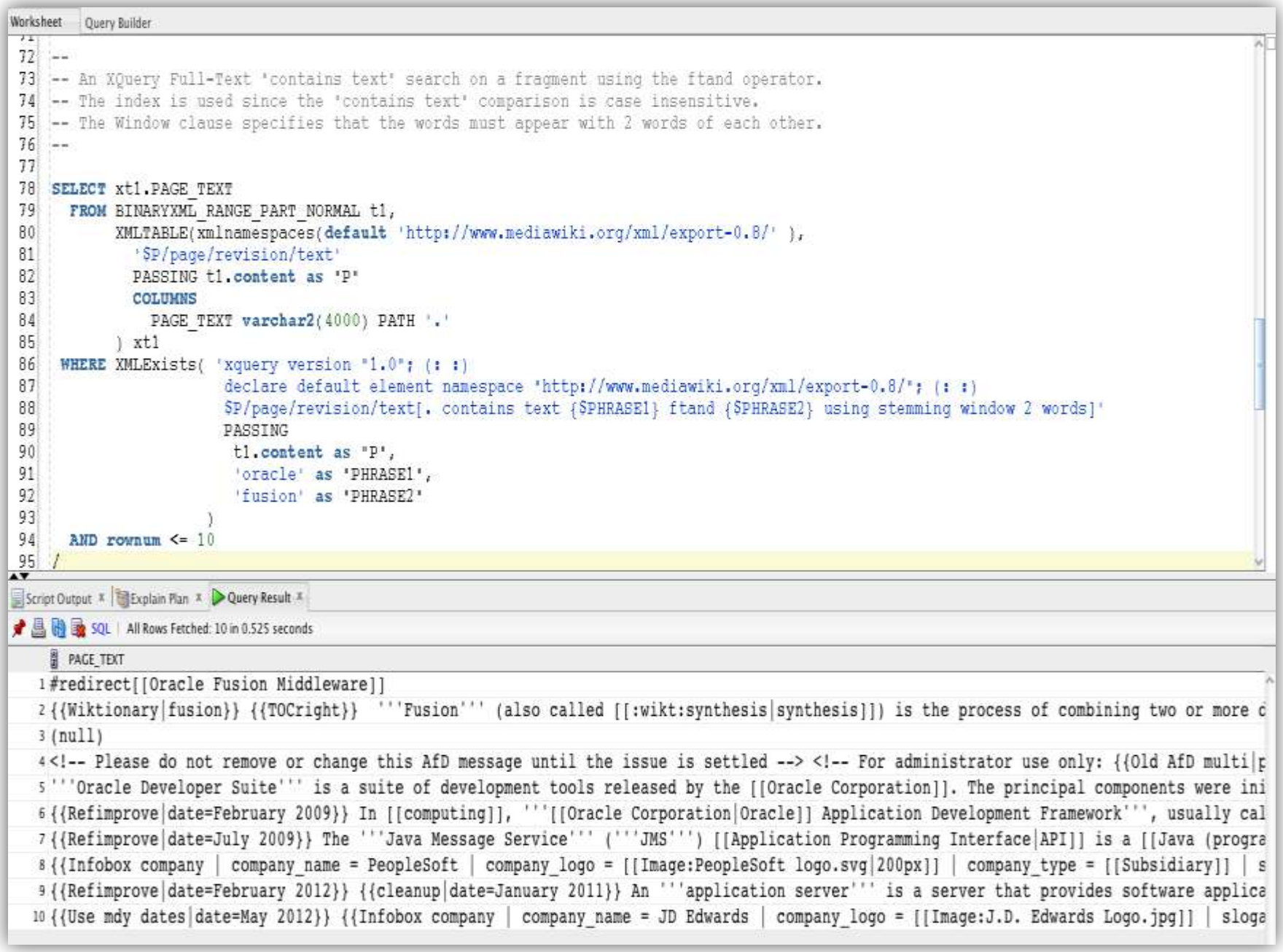
	PAGE_ID	PAGE_TITLE	PAGE_REV_TIMESTAMP
1	1754	Andrea Andreani	09-JUL-12 10.12.29.000000000 AM +00:00

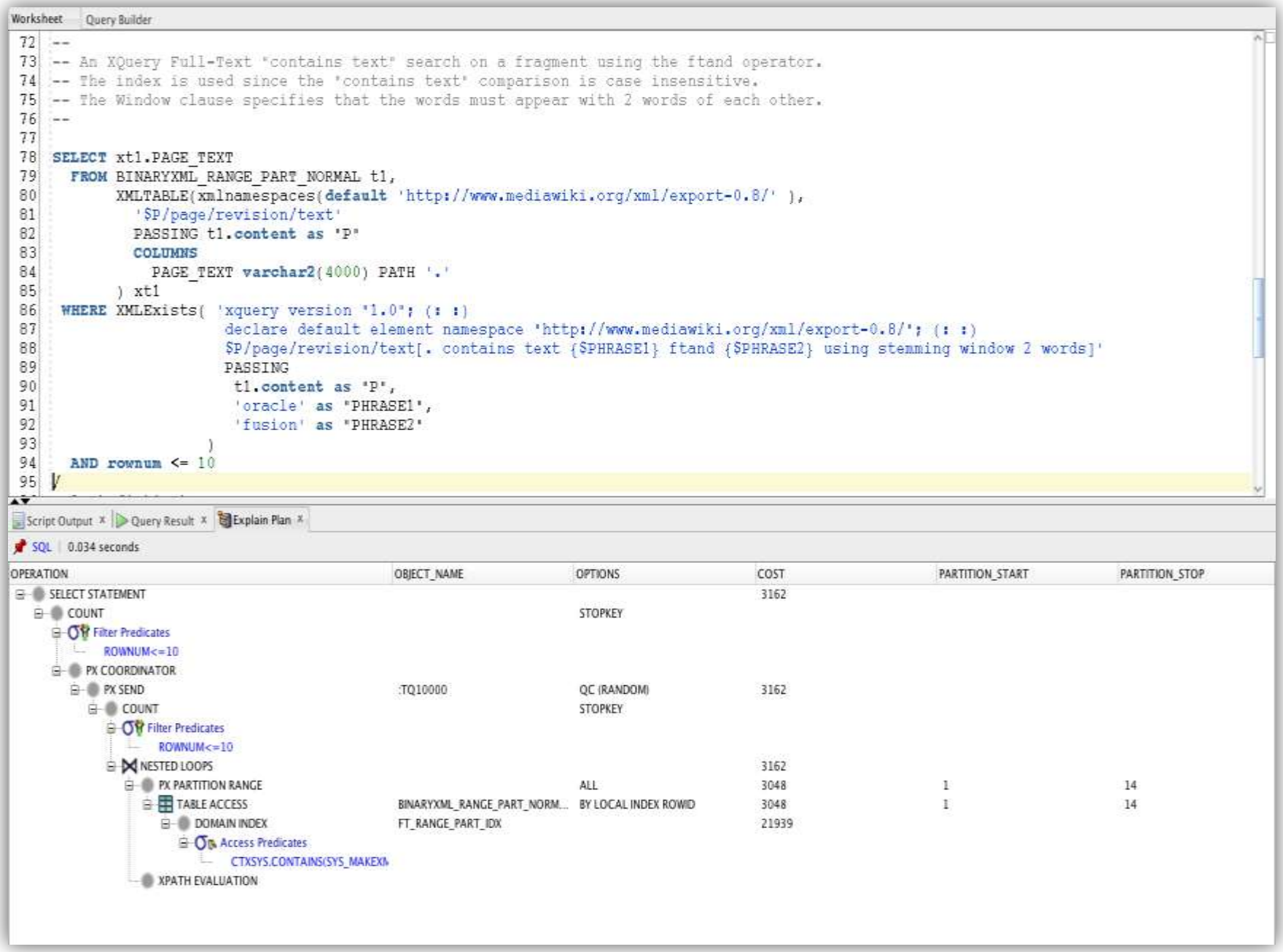
Un-Structured Data

XML Full Tekst Index

- Based on Oracle Text Index, XQuery Full Text
- XML Namespace Aware
- XML Semantic aware full text search
  - Full-Tekst Selection Expression – contains text
  - Logical Full Text Operator – ftor, ftand, ftMildNot
  - Context Aware full text search

```
SELECT po.id
  FROM purchaseorder po
 WHERE XMLEXISTS ('$src/purchaseOrder/billingInstruction/Address
                   [.contains text {$PHRASE1) ftand {$PHRASE2) using stemming]
                   PASSING po.x,
                           'Science'  as "PHRASE1",
                           'Magdalen' as "PHRASE2"
                  )
```


Worksheet
Query Builder
--
-- An XQuery Full-Text 'contains text' search on a fragment using the ftand operator.
-- The index is used since the 'contains text' comparison is case insensitive.
-- The Window clause specifies that the words must appear with 2 words of each other.
--
SELECT xt1.PAGE_TEXT
FROM BINARYXML_RANGE_PART_NORMAL t1,
XMLTABLE(xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/'),
'$P/page/revision/text'
PASSING t1.content as 'P'
COLUMNS
PAGE_TEXT varchar2(4000) PATH '.'
) xt1
WHERE XMLExists('xquery version "1.0"; (: :)
declare default element namespace 'http://www.mediawiki.org/xml/export-0.8/"; (: :)
$P/page/revision/text[. contains text {$PHRASE1} ftand {$PHRASE2} using stemming window 2 words]'
PASSING
t1.content as "P",
'oracle' as 'PHRASE1',
'fusion' as 'PHRASE2'
)
AND rownum <= 10
/
Script Output
Explain Plan
Query Result
SQL
All Rows Fetched: 10 in 0.525 seconds
PAGE_TEXT
#redirect[[Oracle Fusion Middleware]]
{{Wiktionary|fusion}} {{TOCright}} '''Fusion''' (also called [[:wikt:synthesis|synthesis]]) is the process of combining two or more c
(null)
<!-- Please do not remove or change this AfD message until the issue is settled --> <!-- For administrator use only: {{Old AfD multi|p
'''Oracle Developer Suite''' is a suite of development tools released by the [[Oracle Corporation]]. The principal components were ini
{{Refimprove|date=February 2009}} In [[computing]], '''[[Oracle Corporation|Oracle]] Application Development Framework''', usually cal
{{Refimprove|date=July 2009}} The '''Java Message Service''' ('''JMS''') [[Application Programming Interface|API]] is a [[Java (progra
{{Infobox company | company_name = PeopleSoft | company_logo = [[Image:PeopleSoft logo.svg|200px]] | company_type = [[Subsidiary]] | s
{{Refimprove|date=February 2012}} {{cleanup|date=January 2011}} An '''application server''' is a server that provides software applica
{{Use mdy dates|date=May 2012}} {{Infobox company | company_name = JD Edwards | company_logo = [[Image:J.D. Edwards Logo.jpg]] | sloga


Worksheet
Query Builder
--
-- An XQuery Full-Text "contains text" search on a fragment using the ftand operator.
-- The index is used since the 'contains text' comparison is case insensitive.
-- The Window clause specifies that the words must appear with 2 words of each other.
--
SELECT xt1.PAGE_TEXT
FROM BINARYXML_RANGE_PART_NORMAL t1,
XMLTABLE(xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/'),
'$P/page/revision/text'
PASSING t1.content as "P"
COLUMNS
PAGE_TEXT varchar2(4000) PATH '.'
) xt1
WHERE XMLExists('xquery version "1.0"; (: :)
declare default element namespace 'http://www.mediawiki.org/xml/export-0.8/'; (: :)
$P/page/revision/text[. contains text {$PHRASE1} ftand {$PHRASE2} using stemming window 2 words]'
PASSING
t1.content as "P",
'oracle' as "PHRASE1",
'fusion' as "PHRASE2"
)
AND rownum <= 10
Script Output
Query Result
Explain Plan
SQL
0.034 seconds
OPERATION
OBJECT_NAME
OPTIONS
COST
PARTITION_START
PARTITION_STOP
SELECT STATEMENT
3162
COUNT
STOPKEY
Filter Predicates
ROWNUM<=10
PX COORDINATOR
PX SEND
:TQ10000
QC (RANDOM)
3162
COUNT
STOPKEY
Filter Predicates
ROWNUM<=10
NESTED LOOPS
3162
PX PARTITION RANGE
ALL
3048
1
14
TABLE ACCESS
BINARYXML_RANGE_PART_NORM...
BY LOCAL INDEX ROWID
3048
1
14
DOMAIN INDEX
FT_RANGE_PART_IDX
21939
Access Predicates
CTXSYS.CONTAINS(SYS_MAKEXM
XPATH EVALUATION

Worksheet | Query Builder

```
114 --
115 -- Window clause of 2, while searching 1.000.000 wikipedia pages
116 --
117 SELECT count(*)
118   FROM BINARYXML_RANGE_PART_NORMAL t1,
119        XMLTABLE(xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/' ),
120          '$P/page/revision/text'
121          PASSING t1.content as "P"
122          COLUMNS
123            PAGE_TEXT varchar2(4000) PATH '.'
124        ) xt1
125  WHERE XMLExists( 'xquery version "1.0"; (: :)
126                    declare default element namespace "http://www.mediawiki.org/xml/export-0.8/"; (: :)
127                    $P/page/revision/text[. contains text {$PHRASE1} ftand {$PHRASE2} using stemming window 2 words]'
128                    PASSING
129                     t1.content as "P",
130                     'oracle'   as "PHRASE1",
131                     'ellison'  as "PHRASE2"
132                  )
133 /
```

Script Output x | Explain Plan x | Query Result x

SQL | All Rows Fetched: 1 in 0.05 seconds

	COUNT(*)
1	3

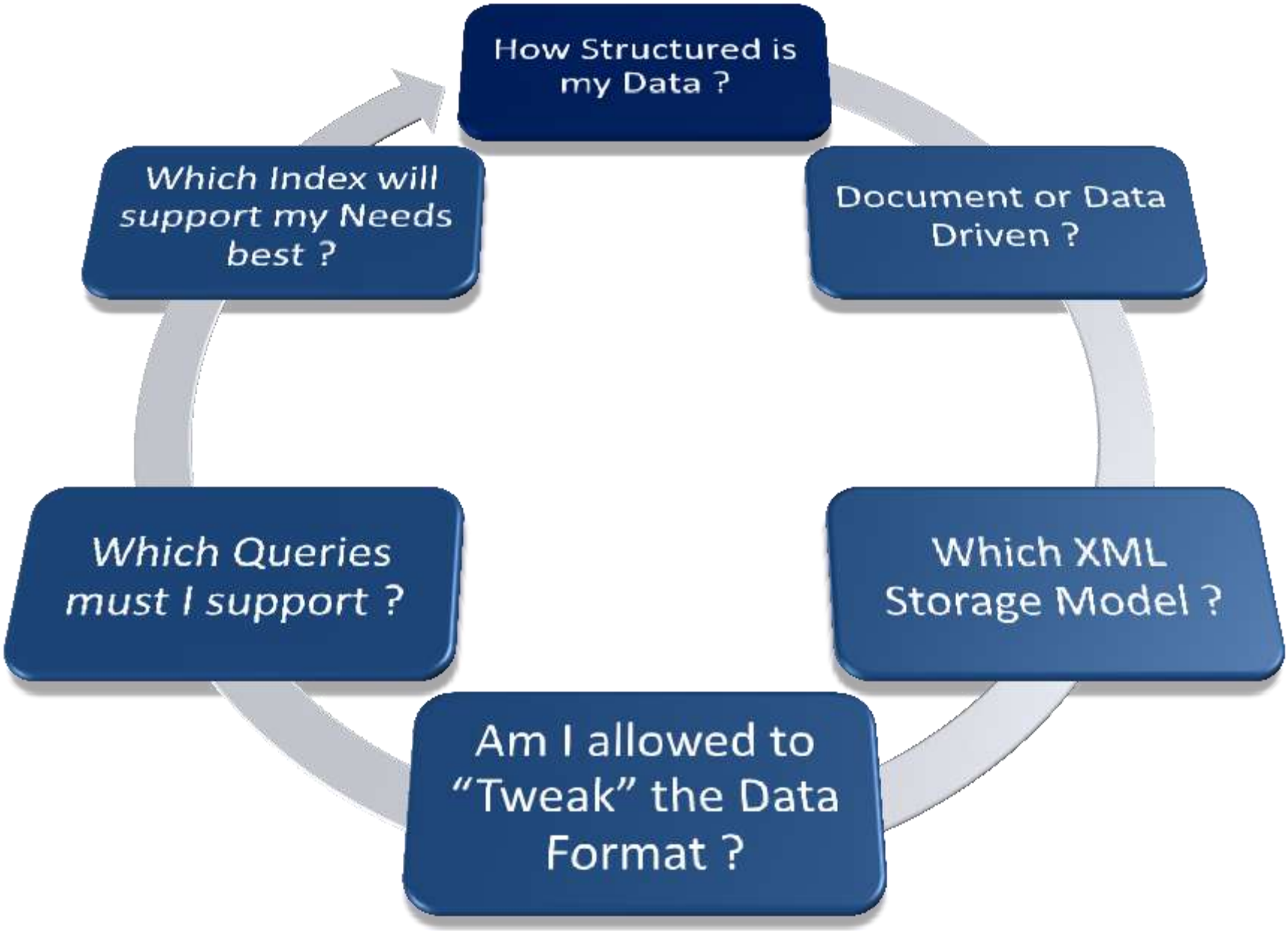
Worksheet | Query Builder

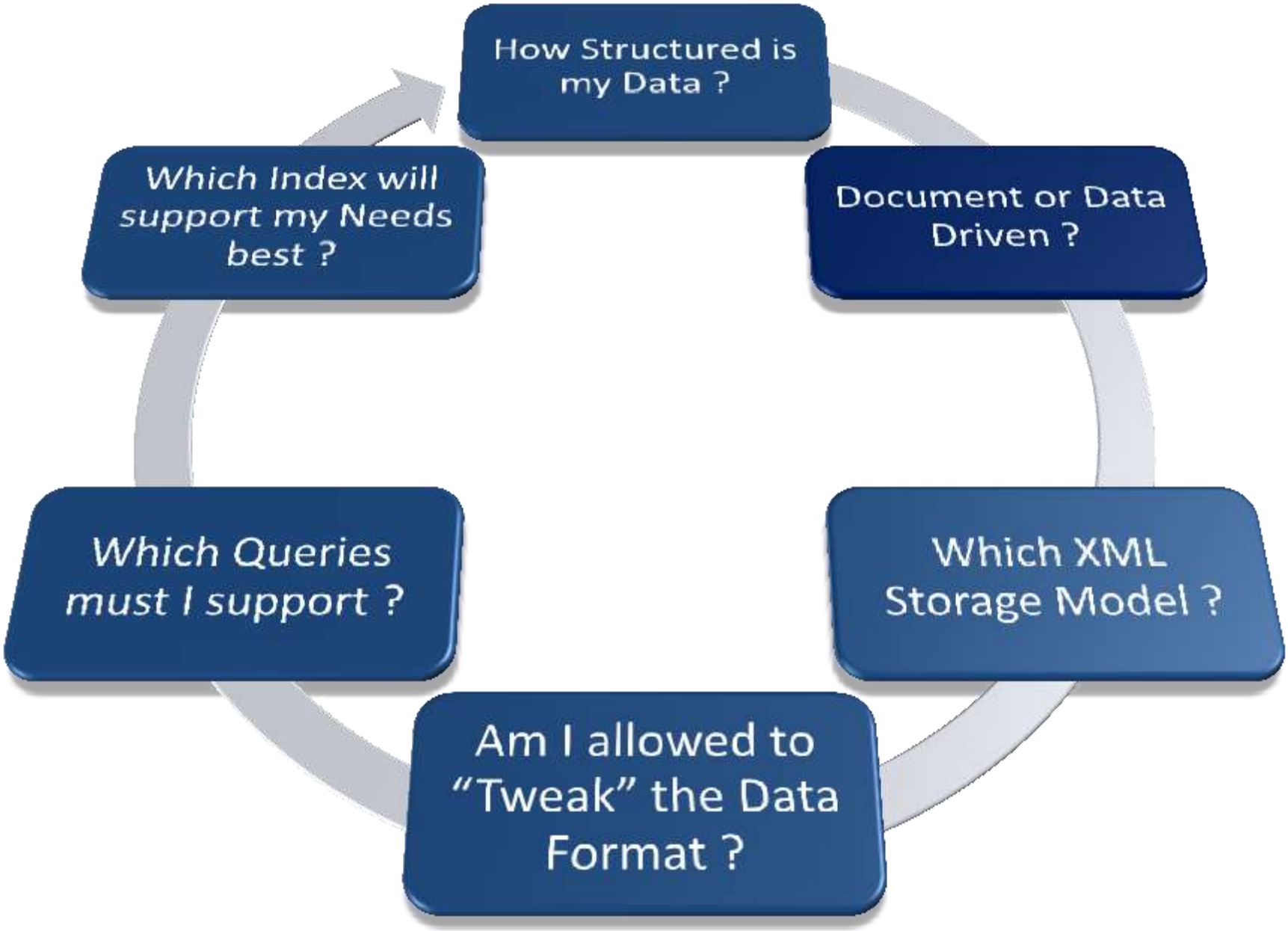
```
134 --
135 -- Window clause of 6, while searching 1.000.000 wikipedia pages
136 --
137 SELECT count(*)
138   FROM BINARYXML_RANGE_PART_NORMAL t1,
139        XMLTABLE(xmlnamespaces(default 'http://www.mediawiki.org/xml/export-0.8/' ),
140          '$P/page/revision/text'
141          PASSING t1.content as "P'
142          COLUMNS
143            PAGE_TEXT varchar2(4000) PATH '.'
144        ) xt1
145  WHERE XMLExists( 'xquery version '1.0'; (: :)
146                    declare default element namespace "http://www.mediawiki.org/xml/export-0.8/"; (: :)
147                    $P/page/revision/text[. contains text {$PHRASE1} ftand {$PHRASE2} using stemming window 6 words]'
148                    PASSING
149                     t1.content as 'P',
150                     'oracle'   as 'PHRASE1',
151                     'ellison'  as 'PHRASE2'
152                  )
153 /
154
```

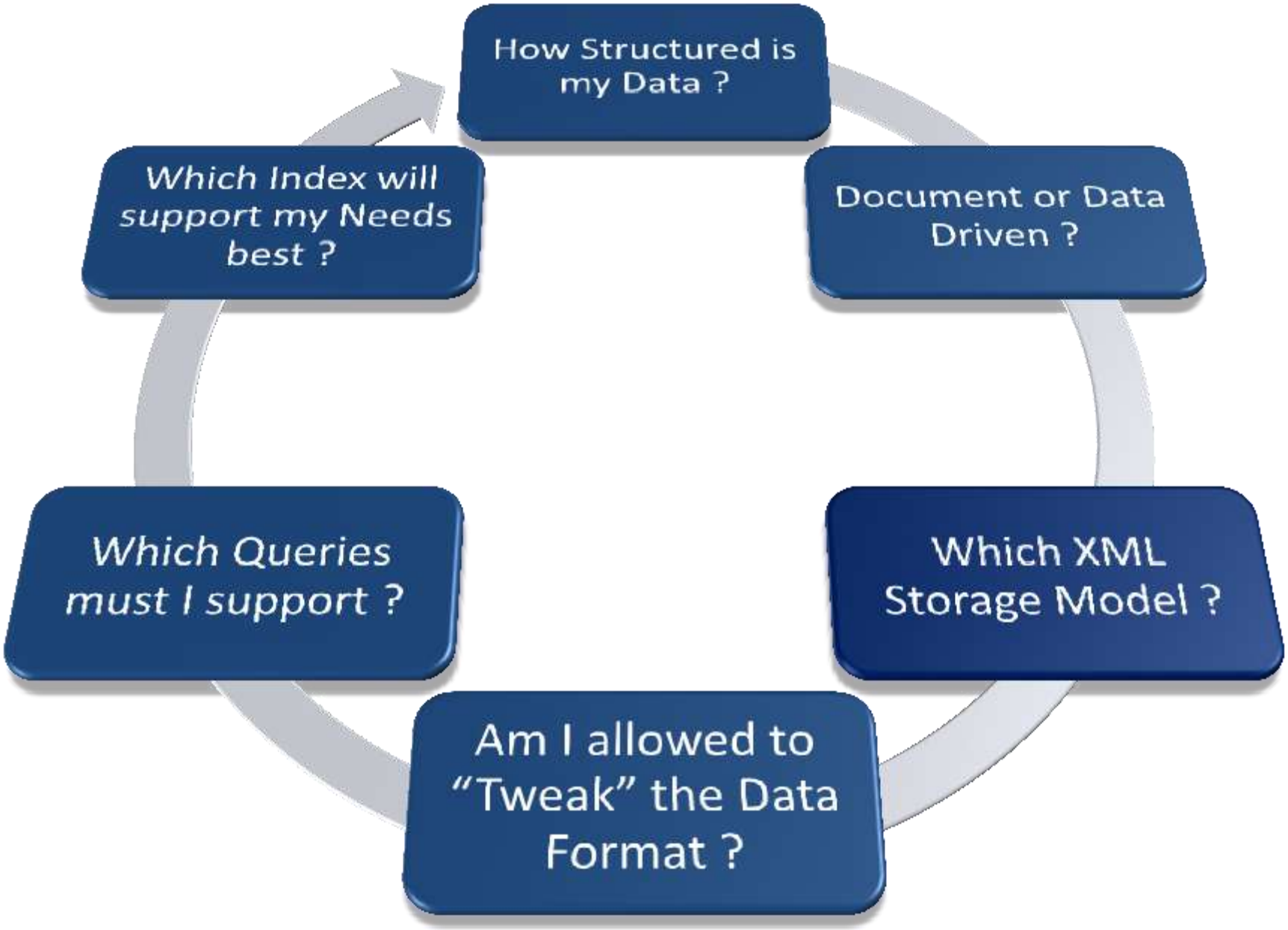
Script Output | Explain Plan | Query Result

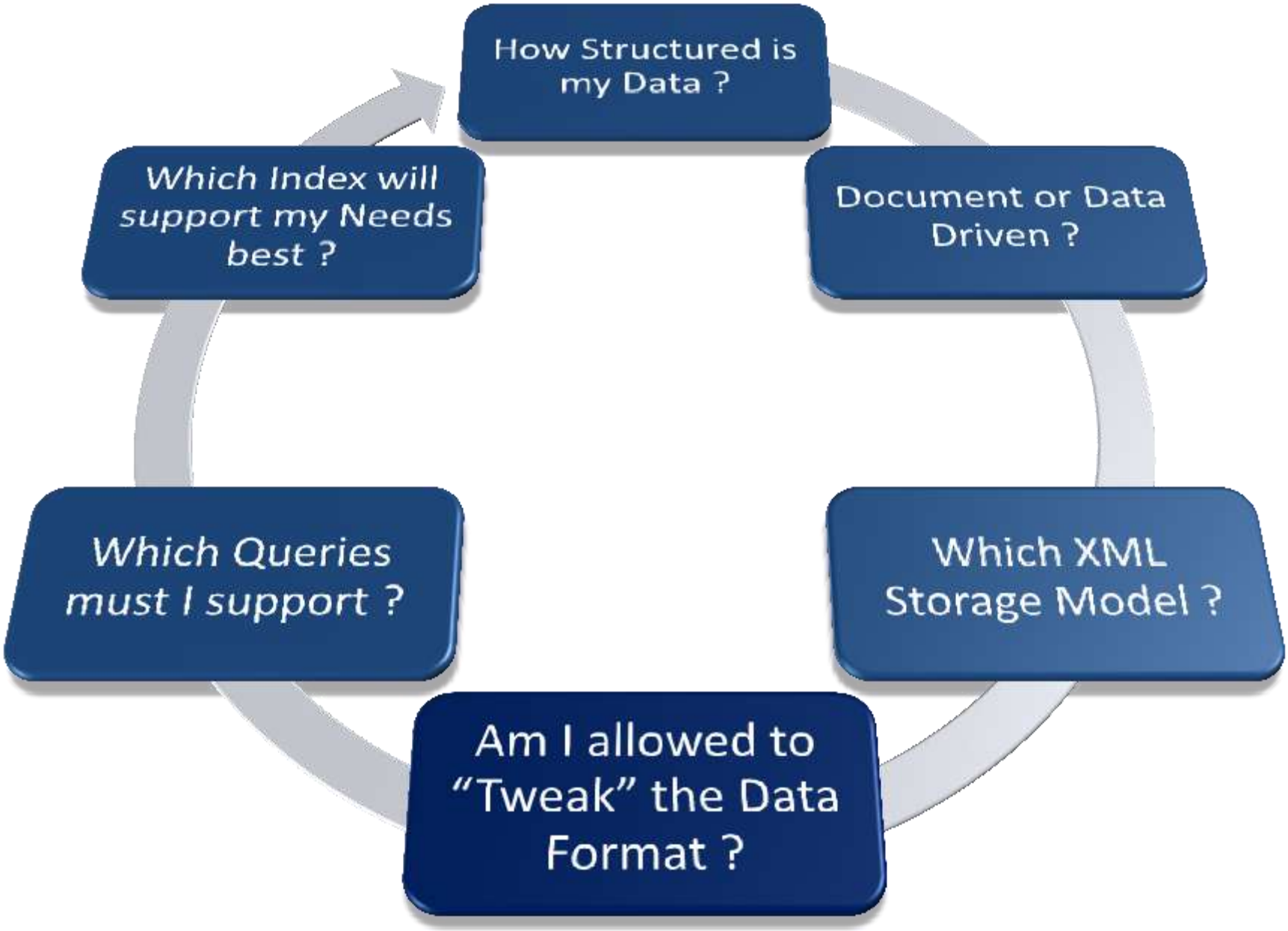
SQL | All Rows Fetched: 1 in 0.054 seconds

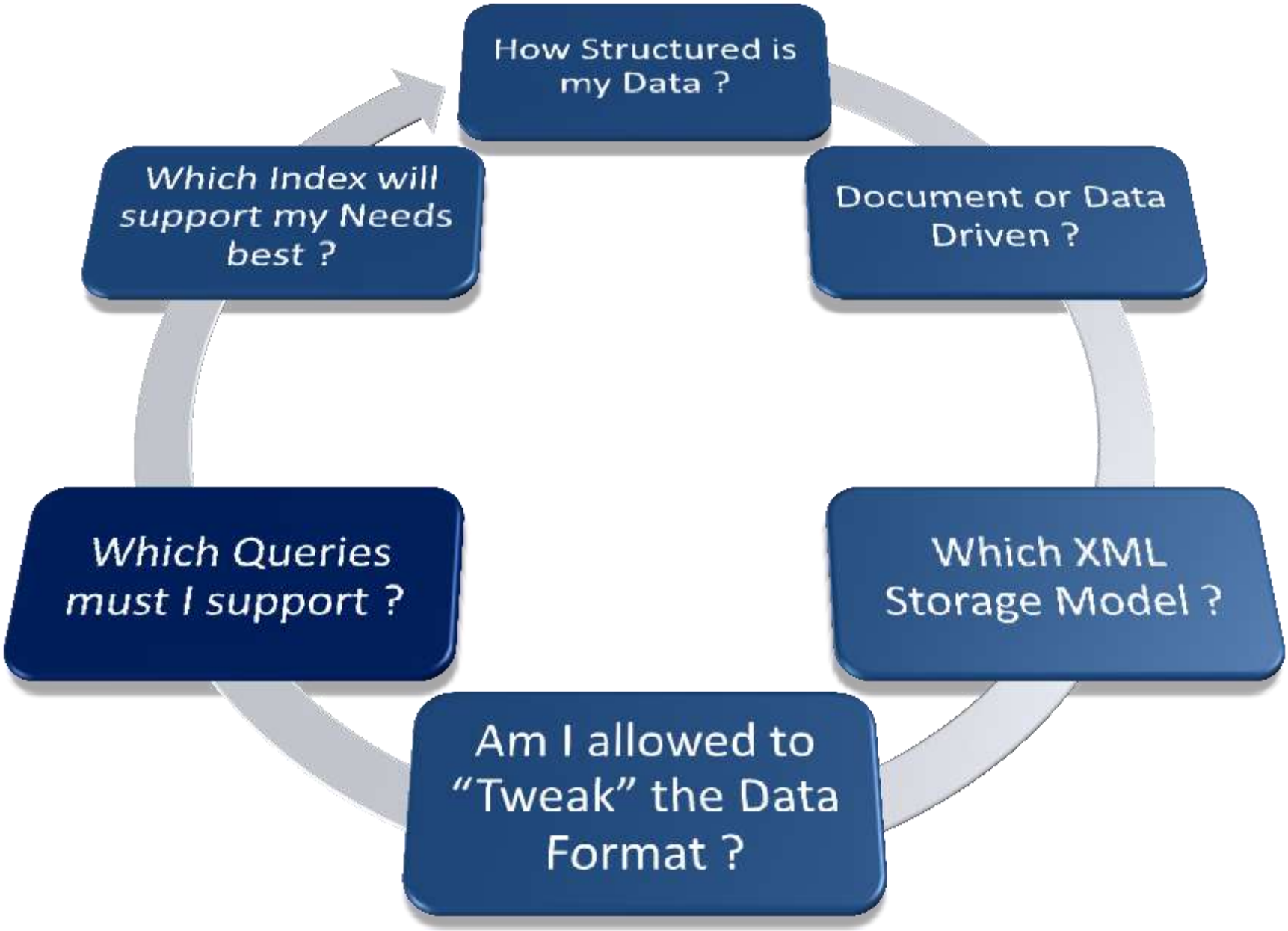
	COUNT(*)
1	32

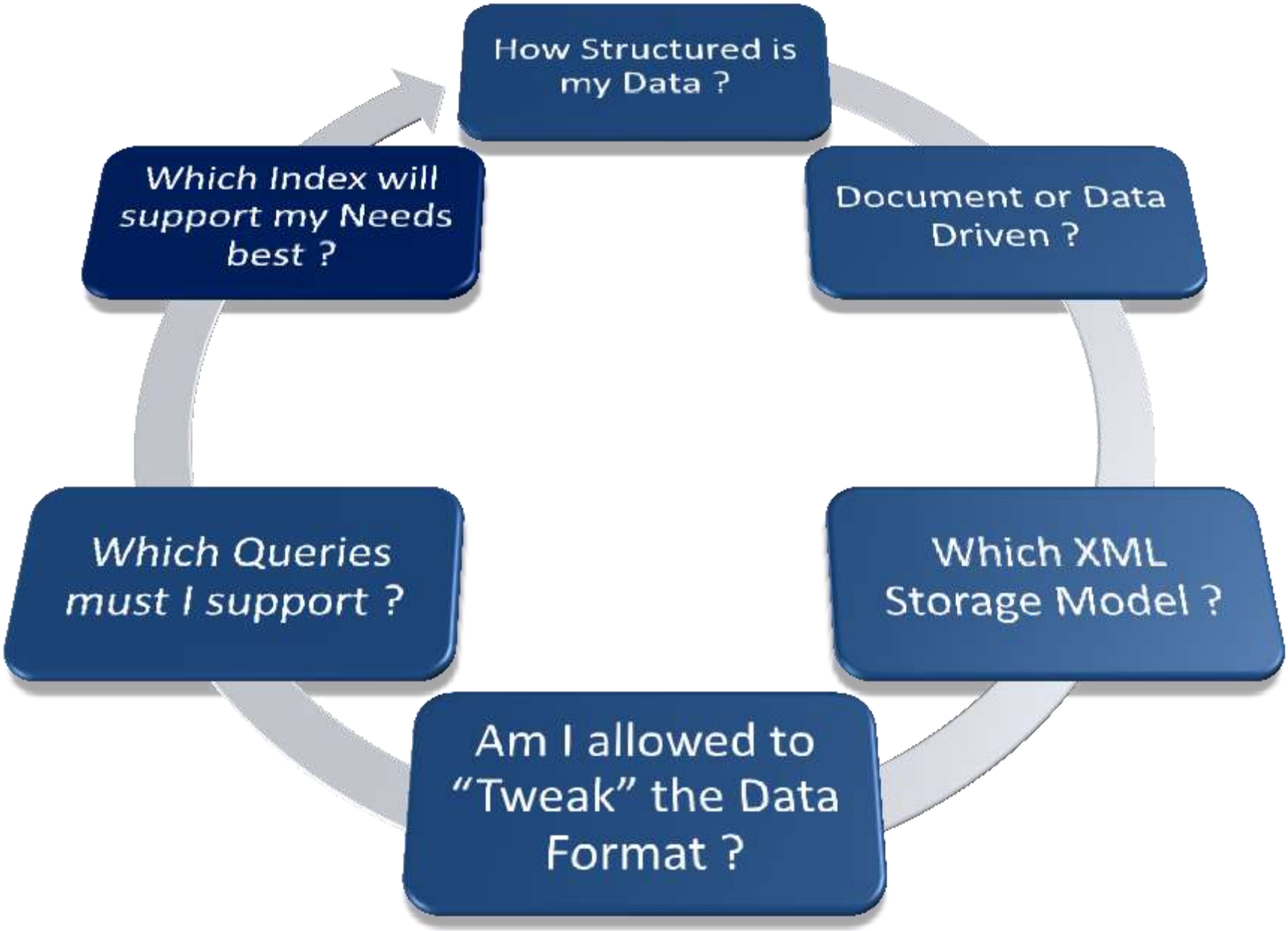

How Structured is my Data ?
Document or Data Driven ?
Which XML Storage Model ?
Am I allowed to "Tweak" the Data Format ?
Which Queries must I support ?
Which Index will support my Needs best ?


How Structured is my Data ?
Document or Data Driven ?
Which XML Storage Model ?
Am I allowed to "Tweak" the Data Format ?
Which Queries must I support ?
Which Index will support my Needs best ?


How Structured is my Data ?
Document or Data Driven ?
Which XML Storage Model ?
Am I allowed to "Tweak" the Data Format ?
Which Queries must I support ?
Which Index will support my Needs best ?

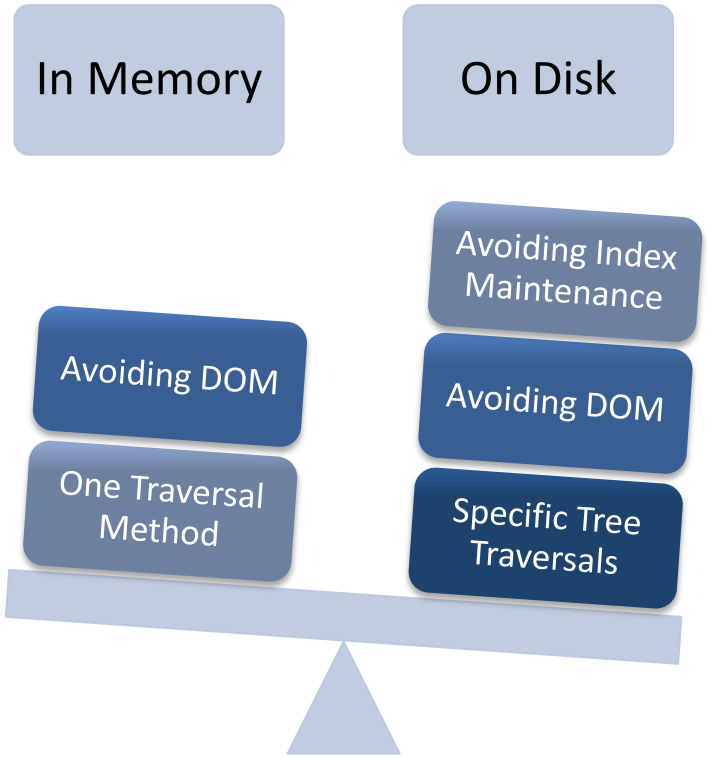

How Structured is my Data ?
Document or Data Driven ?
Which XML Storage Model ?
Am I allowed to "Tweak" the Data Format ?
Which Queries must I support ?
Which Index will support my Needs best ?


How Structured is my Data ?
Document or Data Driven ?
Which XML Storage Model ?
Am I allowed to "Tweak" the Data Format ?
Which Queries must I support ?
Which Index will support my Needs best ?


How Structured is my Data ?
Document or Data Driven ?
Which XML Storage Model ?
Am I allowed to "Tweak" the Data Format ?
Which Queries must I support ?
Which Index will support my Needs best ?

Balanced Design

- Inserts, Updates & Deletes
  - XML Future Changes
  - Index Maintenance
- Selects
  - In Memory
  - Via Indexes
- XML Validation
  - Strict, Lazy
  - Client Side Possibilities



Reward

- Optimal performance
- Out performing XML
- Proper design will give performance increase over XML handling…

…proper design is still key…

References

Oracle XML DB

- http://www.oracle.com/pls/db112/homepage

XML DB FAQ Thread

- http://forums.oracle.com/forums/thread.jspa?threadID=410714

Personal Blog

- http://www.xmldb.nl
- http://technology.amis.nl

References

Daniela Florescu, Oracle Corporation

Advances in XML and XQuery

Sam Idicula, Oracle XML DB Development Team

Binary XML Storage and Query Processing in Oracle

Jinyu Wang, Scott Brewton

Making XML Technology Easier to Use

Joel Spolsky - Joel on Software

Back to Basics

References

Oracle XML DB Main page material

- Oracle XML DB : Best Practices to Get Optimal Performance out of XML Queries (PDF)
- Oracle XML DB : Choosing the Best XMLType Storage Option for Your Use Case (PDF)
- A Request for Comments for the Oracle Binary XML Format