



Technical Aspects of SIARD



“SIARD under the hood”

10. April 2003 / Stephan Heuscher

Overview



- ⌘ Introduction
- ⌘ Addressing the problems
 - ☑ SQL “standard”?
 - ☑ Character encodings
 - ☑ Metadata in databases
- ⌘ Solutions in SIARD
- ⌘ Conclusion





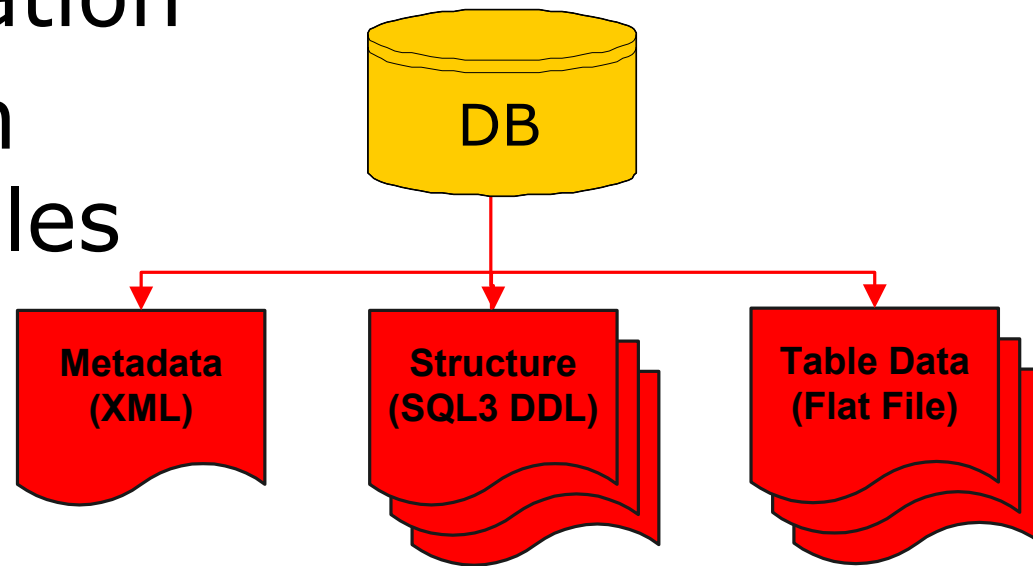
Introduction

SIARD

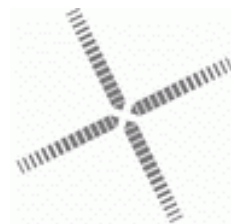
⌘ is a Java application

⌘ archives a DB in three types of files

- ☑ Metadata
- ☑ Structure data
- ☑ Tables



SIARD: Java Application



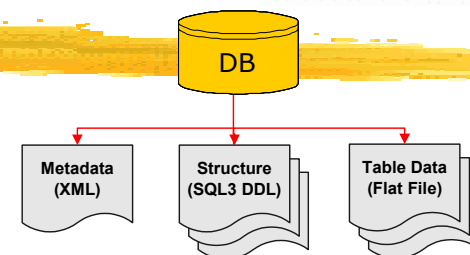
⌘ Platform independence

☑ Write once, run everywhere

⌘ Generic database interface (JDBC)

☑ Drivers available for most database products

⌘ Implemented by Trivadis AG (Switzerland)



SQL “Standard”?

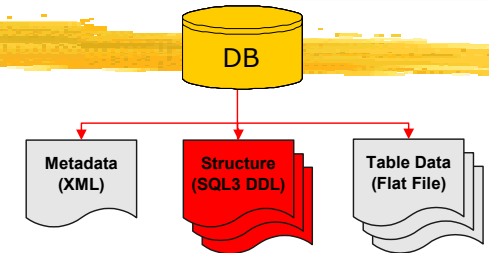


Yes

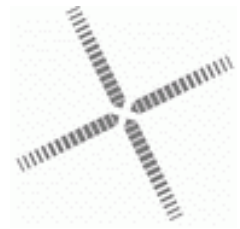
- ⌘ ISO/IEC 9075:1999
- ⌘ Almost every RDBMS is based on SQL

No

- ⌘ Vendor-specific extensions
- ⌘ Only partial conformance
- ⌘ Erroneous implementations



Data Types in Databases



⌘ "Simple"

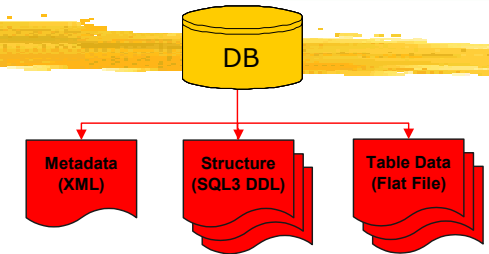
- ☑ Bit, Text, Number, etc.
- ☑ Defined by SQL-3 core

⌘ Complex and complicated

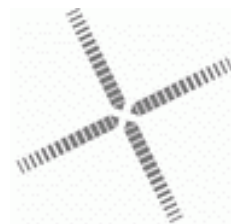
- ☑ Vendor-specific or user-defined Types
- ☑ Must be mapped to SQL-3 core defined types where possible or ...

⌘ Big

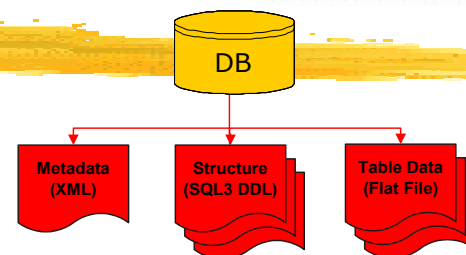
- ☑ Large object (LOB): Character or Binary
- ☑ No standard for storage and retrieval



Character Encodings



- ⌘ Text is not just 7-bit ASCII!
- ⌘ Many possibilities to encode characters in 8 bits (Latin-1 to 15, Code pages, etc)
- ⌘ Different encodings are in use because of multilingualism



Low-level Metadata



⌘ Data types

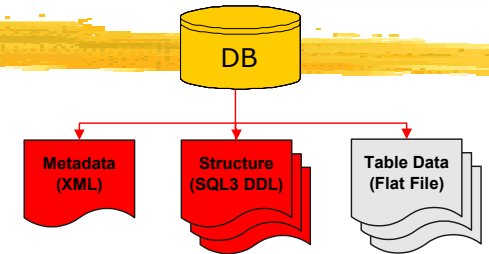
⌘ Structure

☑ Basic: Schemas, Tables and Views, Columns

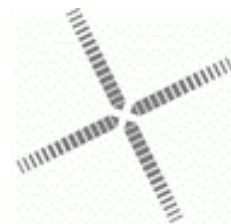
☑ Integrity: Primary Keys, Foreign Keys, Constraints...

⌘ Management and operational data

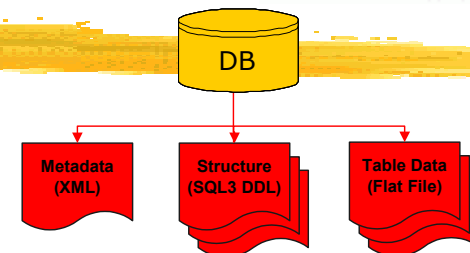
☑ Users, Rights, Indexes, Logs...



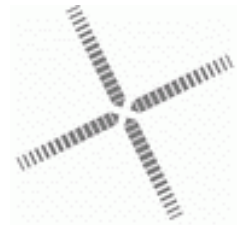
Solution: The Basics



- ⌘ All generated files are text
- ⌘ Structure and data types are represented using the SQL-3 DDL
 - ☑ Clearly defined data types
 - ☑ Automatic lossless conversion to SQL-3 types
 - ☑ Parser checks SQL-3 compliance of extracted statements
- ⌘ Metadata is stored in XML
 - ☑ Semantic markup
 - ☑ High flexibility built-in
 - ☑ Clear exit strategy



Metadata Structure

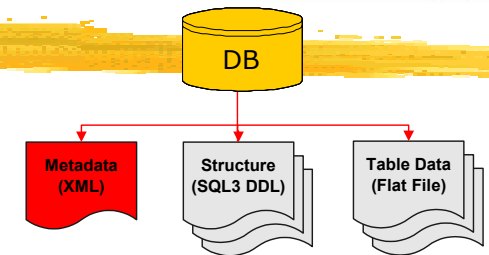


⌘ High-level metadata

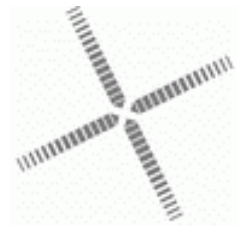
- ☑ Human generated
- ☑ Documents usage, provenance, aims, development, etc of the database
- ☑ Highly flexible to suit future needs
- ☑ Defined externally to fit the users' needs

⌘ Low-level metadata

- ☑ Automatically generated from database contents, structure and technical data



SIARD XML Metadata



```
<?xml version="1.0" encoding="UTF-16"?>
```

```
<archive a0-version="0.8" a1-version="0.8"  
  access-mode="ch.admin.areda.siard.a0.mode.Oracle8iMode">
```

```
<usage>
```

```
<document-list />
```

```
<generic tag-name="System ID">
```

```
<generic tag-name="System Name">
```

```
<attribute key="shortname" mandatory="yes" value="AMDA" description="Name under which  
this database or system was most commonly referred to." />
```

```
<attribute key="longname" mandatory="yes" value="Audio Meta Data Acquisition"  
description="Full name of this database or system." />
```

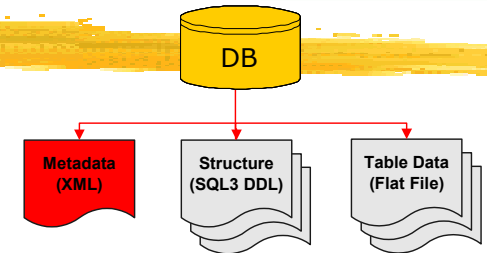
```
<!-- ... -->
```

```
</usage>
```

```
<database product-name="Oracle" product-version="Personal Oracle9i Release 9.0.1.1.1 -  
Production. With the Partitioning option. JServer Release 9.0.1.1.1 - Production" table-  
number="22" view-number="4" archive-size="175KB">
```

```
<schemas>
```

```
<!-- ... -->
```



Archive Data Storage

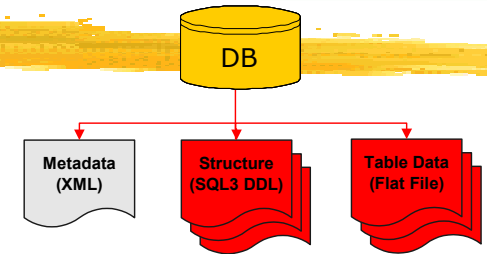


⌘ All files stored in UTF-16

- ☑ Oldest form of encoding Unicode
- ☑ Unicode: All characters, extremely stable
- ☑ Balances storage, applicability and reusability
- ☑ 1:1 mapping for common characters (One 16-bit sequence per character)

⌘ Structure stored in SQL-3 DDL

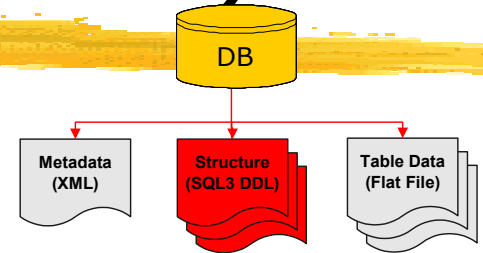
- ☑ As defined by the SQL-3 Standard



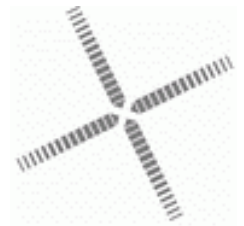
Data Description (Structure)

```
CREATE TABLE "FLUGLE"."CLASS"  
( "CLASS_ID" NATIONAL CHARACTER VARYING(20) NOT NULL  
, "SCHEDULE_ID" NATIONAL CHARACTER VARYING(20)  
, "CLASS_BUILDING" NATIONAL CHARACTER VARYING(25)  
, "CLASS_ROOM" NATIONAL CHARACTER VARYING(25)  
, "COURSE_ID" NATIONAL CHARACTER VARYING(5)  
, "DEPARTMENT_ID" NATIONAL CHARACTER VARYING(20)  
, "INSTRUCTOR_ID" NATIONAL CHARACTER VARYING(20)  
, "SEMESTER" NATIONAL CHARACTER VARYING(6)  
, "SCHOOL_YEAR" TIMESTAMP(0)  
)
```

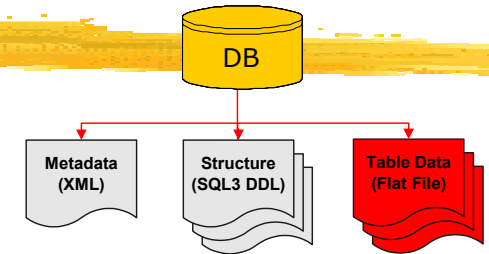
```
CREATE TABLE "FLUGLE"."CLASS_LOCATION"  
( "CLASS_BUILDING" NATIONAL CHARACTER VARYING(25) NOT NULL  
, "CLASS_ROOM" NATIONAL CHARACTER VARYING(25) NOT NULL  
...
```



SIARD XML-ized Data



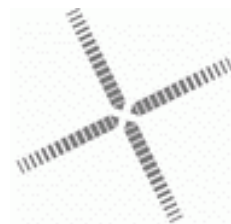
```
<?xml version="1.0" encoding="UTF-16"?>
<dmp-file xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="../dmp.xsd">
  <schema tag-name="FLUGLE"/>
  <table tag-name="CLASS"/>
  <column tag-name="CLASS_ID" sql3type="NATIONAL CHARACTER VARYING"
sql3size="(20)" defaultvalue="" nullable="false" constraints="PK:PK_CLASS"/>
<!-- ...
```



↑ Added value ↑
↓ Data ↓

```
-->
  <data>
<row>6,104200;4,S180;9,POCO HALL;3,150;3,198;5,PHILO;4,E491;6,SPRING;19,1997-03-01
00:00:00;</row>
  <row>6,104500;3,T15;11,NARROW HALL;3,200;3,184;4,HIST;4,D944;6,SPRING;19,1997-
03-01 00:00:00;</row>
  <!-- ... -->
```

Conclusion (what we have)



- ⌘ Platform independent set of tools to archive and restore databases
- ⌘ Flexible high-level metadata definition and acquisition
- ⌘ Database in a stable, authentic, easily migratable and accessible form
 - ☑ Text only files
 - ☑ Tested for compliance with completely documented standards
 - ☒ SQL-3, XML 1.0, Unicode 3/UTF-16



Do You want SIARD?



⌘ It's not yet available to the public

- ☒ Beta testing phase

- ☒ Improve workflow and usability for non-archive users (pre-ingest)

⌘ Roadmap

- ☒ End of beta testing: Autumn 2003

- ☒ Interested? Send an email to

stephan.heuscher@bar.admin.ch

