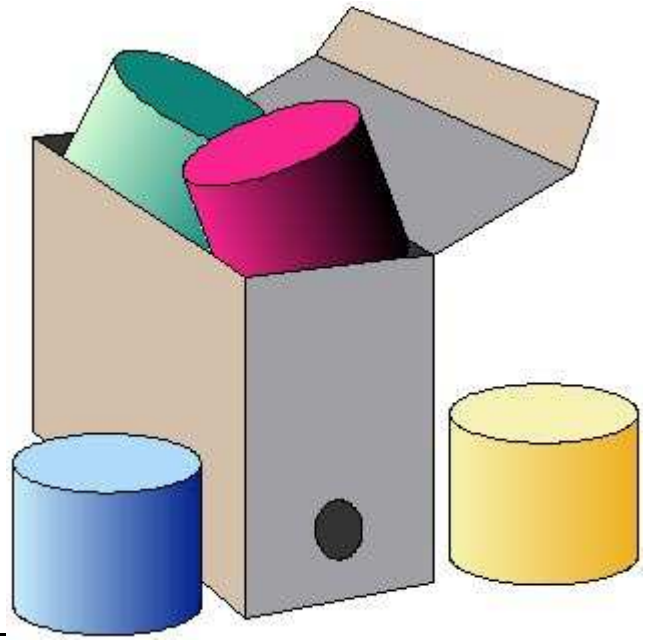
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Schweizerisches Bundesarchiv BAR
Ressort Innovation und Erhaltung

SIARD
A File Format for Archiving Databases

Krystyna W. Ohnesorge, PhD
Hartwig Thomas, PhD
Amir Bernstein, PhD
Anne-Louise Joël, MA



Agenda

- Part 1: Digital Archiving – The Response of the SFA
  By Krystyna W. Ohnesorge, PhD

- Part 2: An Introduction to Archiving Databases with SIARD
  By Hartwig Thomas, PhD

- Part 3: Archiving with SIARD 2.0 – A Presentation of the SIARD Format and Application
  By Amir Bernstein, PhD

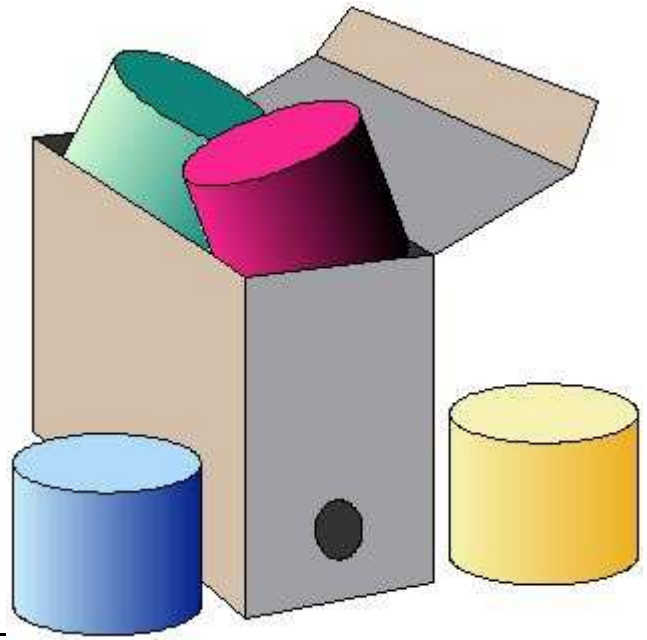
- Part 4: Defining a Database Archiving Process
  By Anne-Louise Joël, MA

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Schweizerisches Bundesarchiv BAR
Ressort Innovation und Erhaltung

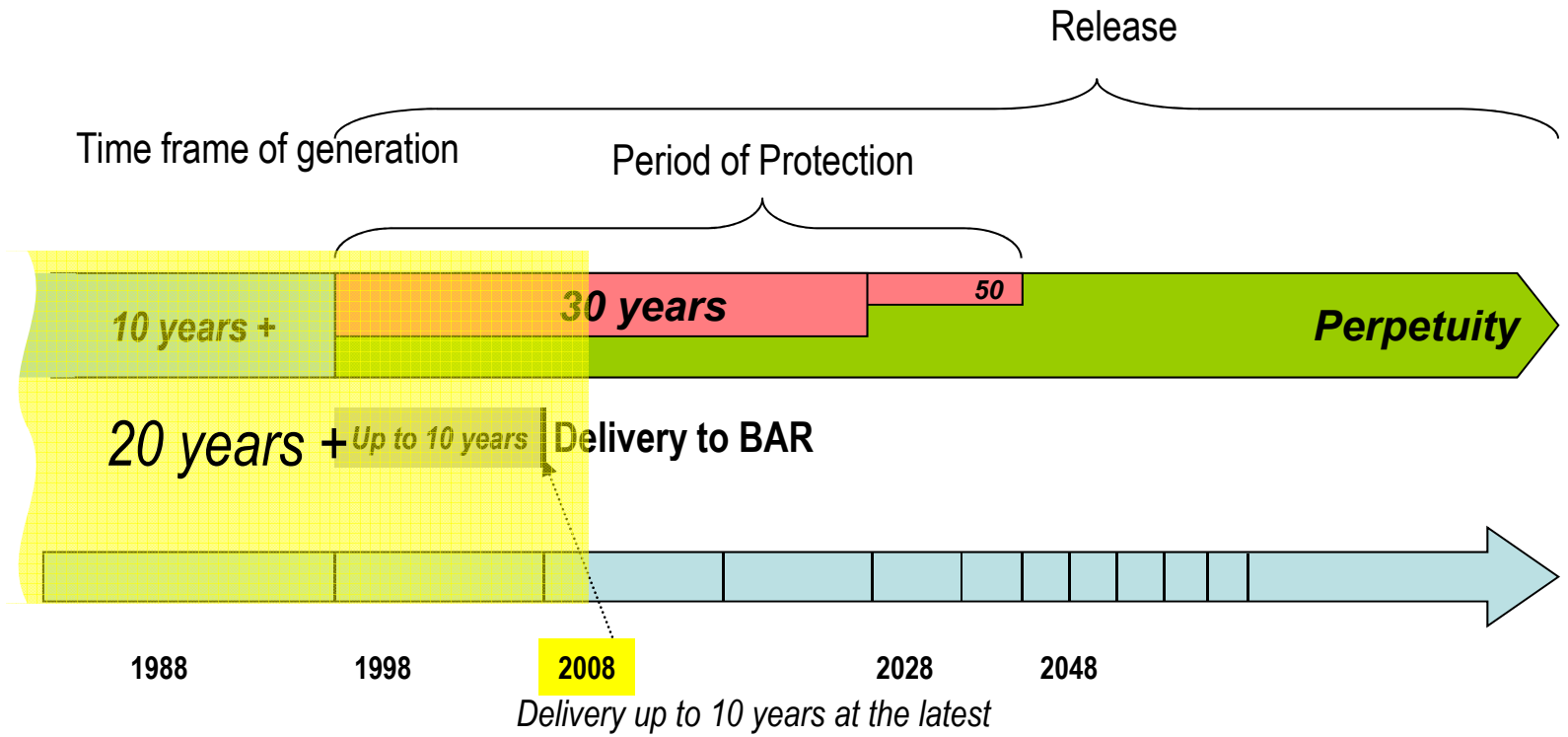
Part 1
Digital Archiving –
The Response of the SFA

A Presentation of the Swiss Federal Archives and the ARELDA Project



Krystyna W. Ohnesorge, PhD

Data Life Cycle

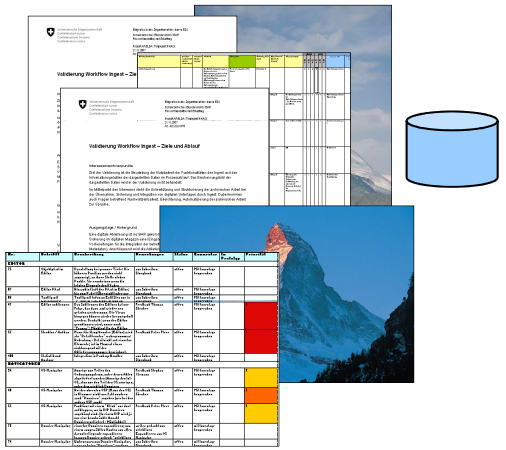


E-Archive – Statistics (end of 2007)

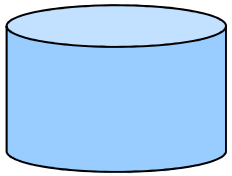
- Number of digital deliveries : ~ 230
- Total of all deliveries:
  - Size in TB: ~ 11.6 TB
  - Number of files: 1.65 Million
- Size of deliveries:
  - smallest - largest 122 KB - 3.5 TB
  - average - median 50 GB - 15 MB
- Number of files per delivery:
  - smallest - largest 1 - 400’000
  - average - median 7’000 - 400

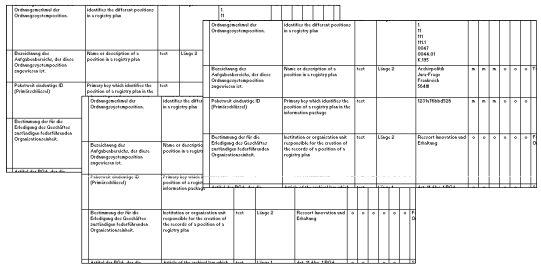
Digital Archive Holdings

Digital dossier



Database





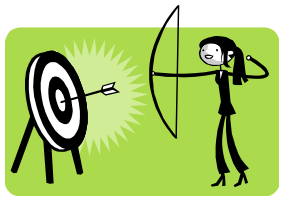
File collections

Strategic Project ARELDA

- First phase: 2001 – 2004:
  - Exploration & experimentation, basic groundwork (**prototype SIARD**: **S**oftware-**I**nvariant **A**rchiving of **R**elational **D**atabases)

- Second phase: 2005 – 2008:
  - Archiving process: ***End-to-End*** Support (from initial advice and consultation, through appraisal, description and submission, to the dissemination of digital records)
  - Integrative approach: organizational and technical aspects (**SIARD 2.0**)

ARELDA – Specifications



- Main focus on the following types of records:

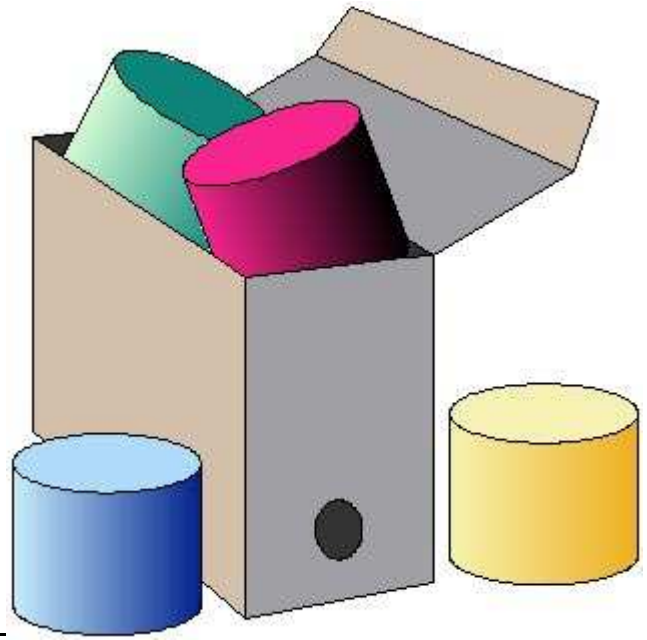
  1st priority: ***Digital Dossiers***
  Documents from records management systems
  (GEVER: Electronic Records Management)

  2nd priority: ***Data from Relational Databases***
  Implementation of SIARD tools and processes
  (**S**oftware-**I**nvariant **A**rchiving of **R**elational **D**atabases)

  3rd priority: ***File Collections***
  Files and their index structure

Part 2
An Introduction to Archiving of Relational Databases using SIARD

Hartwig Thomas, PhD



Introduction: Why data formats matter

Know the alphabet and translate

„Shadrach gave 1 bushel of barley to the temple...“

Try to read the disk by an ancient machine

...10010100100...

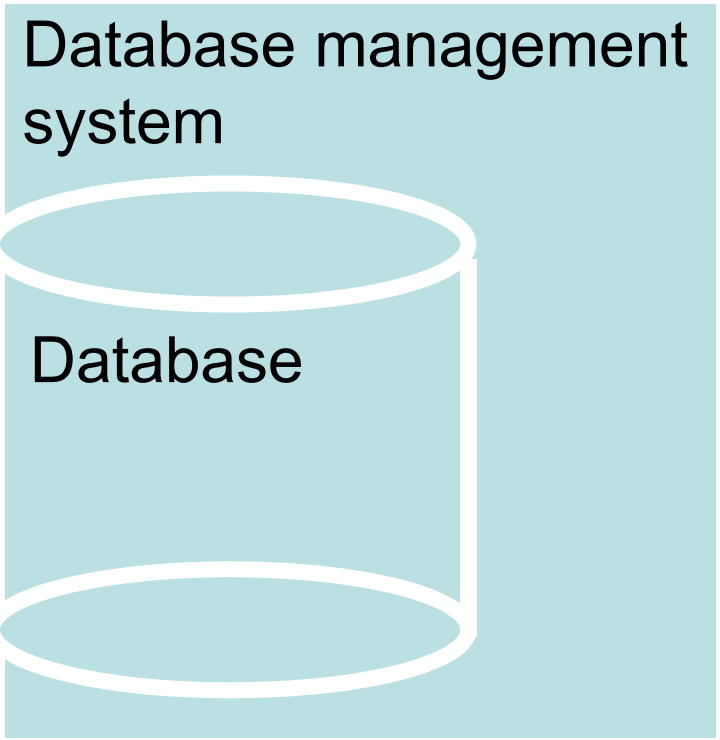
Know the alphabet and translate

...23,010273,9300,00005…

See that it’s a data base. Know the language of that data base. Perform some statements in this language

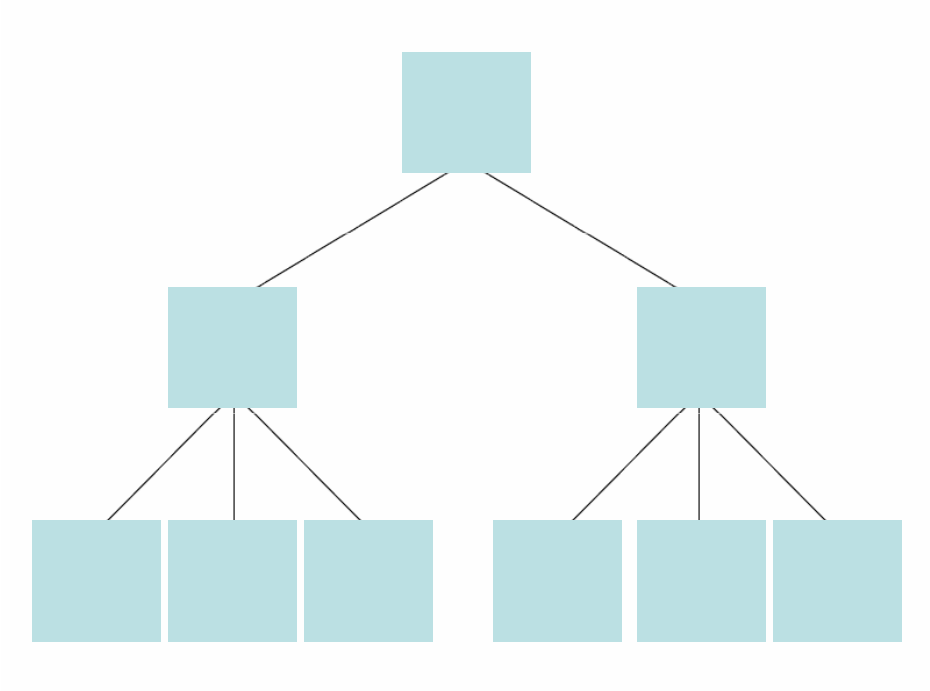
„At the cbot February 1973, the trade limit for barley $0.05 per bushel …

Databases: Basics

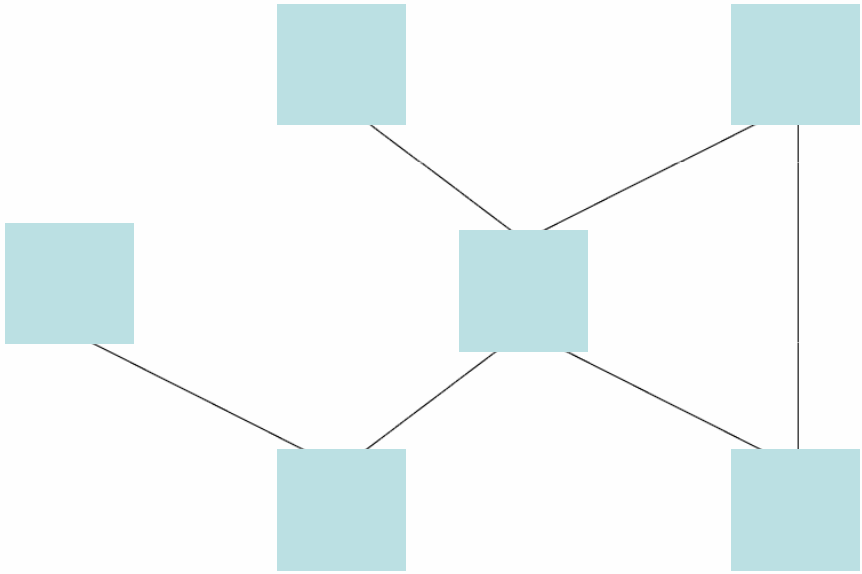


- A database can be described as a repository for a collection of computerized data files
- A database system consists of its data, hardware, software and users

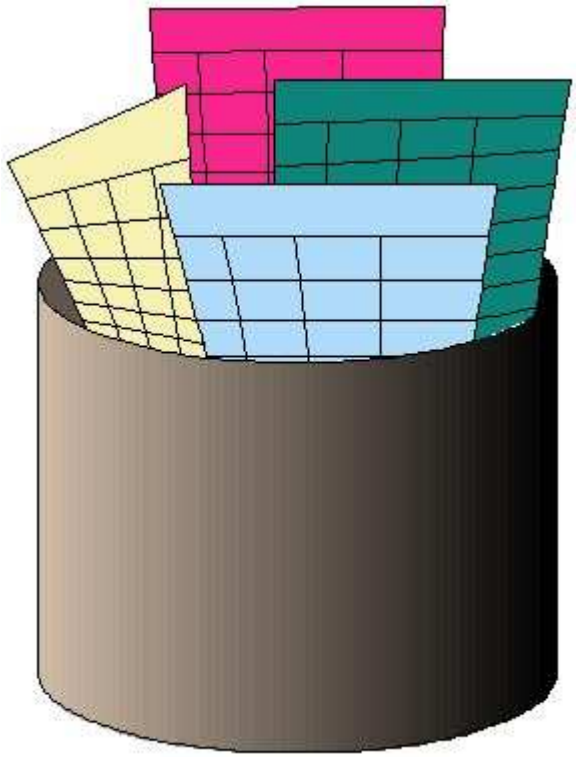
Short History: Hierarchical Data Bases



Short History: Network Data Bases



The Relational Model



Relation is a mathematical term designating something like a *table*, and thus *relational* roughly means "based on tables"

Predicate logic is quite good at describing the state of the world (or of a model of the world), so relational databases must be also quite good at describing the state of the world

The Relational Model

The Relational Model was introduced by Edgar F. Codd around 1970. It remedies some of the defects of the hierarchical and network model.

It is based on the following assumptions:

- Data have a longer life than software, hardware or systems
- Data must be independent of software, hardware or systems
- Changes in the organization of data must be hidden to users
- A user-friendly query language must be standardized
- All queries must be treated equally

The Relational Model

- The Relational Model is conceptual
- The model disconnects the schema (logical organization) of a database from the physical storage methods
  - This is very important for archiving, because it allows the separation of content and media
- The model is concerned with:
  - data structure
  - data integrity
  - data manipulation

Relational Model: ANSI-SPARC 1978

External Level

User defined views

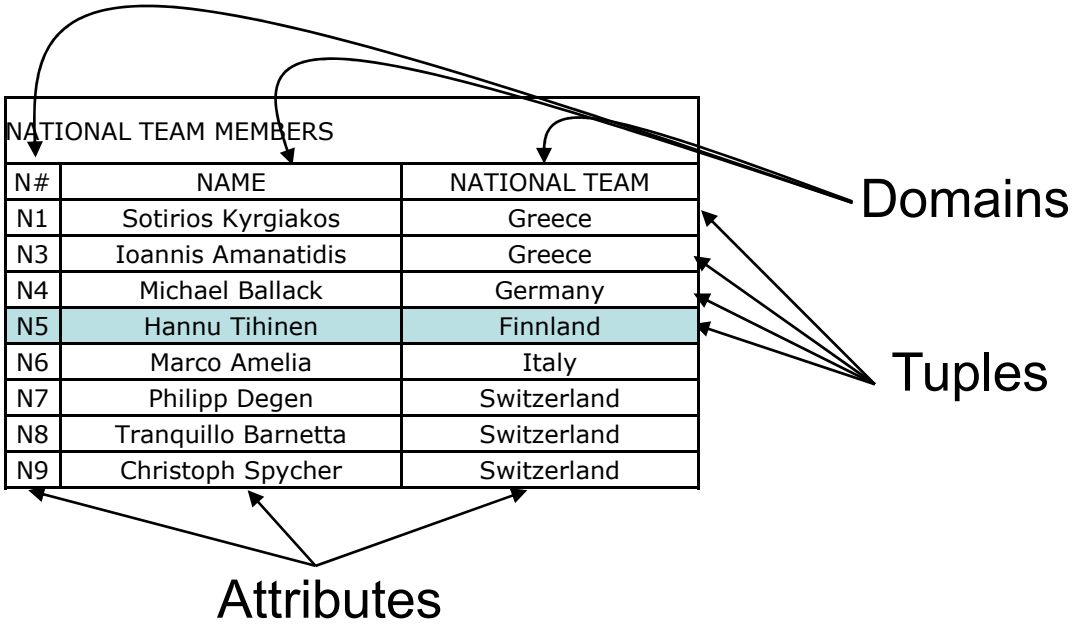
Conceptual Level

Logical view, „community user view“

Internal Level

Physical description (blocks and pages), storage view

The Relational Model



The Relational Model

The Relational Model

NATIONAL

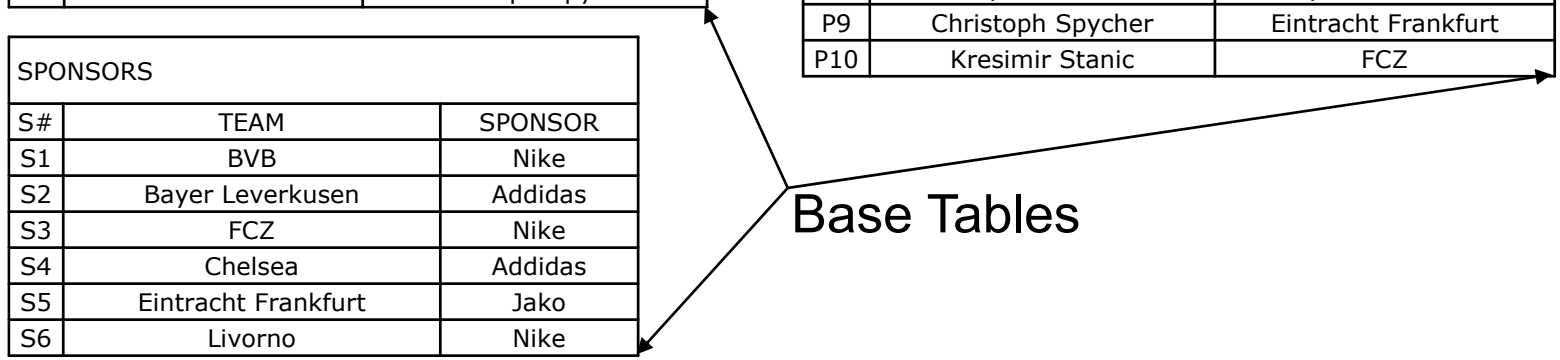
N#	TEAM	PLAYER
N1	Greece	Sotirios Kyrgiakos
N3	Greece	Ioannis Amanatidis
N4	Germany	Michael Ballack
N5	Finnland	Hannu Tihinen
N6	Italy	Marco Amelia
N7	Switzerland	Philipp Degen
N8	Switzerland	Tranquillo Barnetta
N9	Switzerland	Christoph Spycher

LEAGUE

P#	PLAYER	TEAM
P1	Philipp Degen	BVB
P2	Pirmin Schwegler	Bayer Leverkusen
P3	Hannu Tihinen	FCZ
P4	Michael Ballack	Chelsea
P5	Ioannis Amanatidis	Eintracht Frankfurt
P6	Marco Amelia	Livorno
P7	Sotirios Kyrgiakos	Eintracht Frankfurt
P8	Tranquillo Barnetta	Bayer Leverkusen
P9	Christoph Spycher	Eintracht Frankfurt
P10	Kresimir Stanic	FCZ

SPONSORS

S#	TEAM	SPONSOR
S1	BVB	Nike
S2	Bayer Leverkusen	Addidas
S3	FCZ	Nike
S4	Chelsea	Addidas
S5	Eintracht Frankfurt	Jako
S6	Livorno	Nike



The Relational Model

```
SELECT NATIONAL.PLAYER,
NATIONAL.TEAM AS "NATIONAL TEAM",
LEAGUE.TEAM as "LEAGUE TEAM"
FROM NATIONAL, LEAGUE
WHERE LEAGUE.PLAYER =
NATIONAL.PLAYER;
```

PN

PN#	PLAYER	NATIONAL TEAM	LEAGUE TEAM
PN1	Philipp Degen	Switzerland	BVB
PN2	Hannu Tihinen	Finnland	FCZ
PN3	Michael Ballack	Germany	Chelsea
PN4	Ioannis Amanatidis	Greece	Eintracht Frankfurt
PN5	Michael Ballack	Germany	Chelsea
PN6	Marco Amelia	Italy	Livorno
PN7	Sotirios Kyrgiakos	Greece	Eintracht Frankfurt
PN8	Tranquillo Barnetta	Switzerland	Bayer Leverkusen
PN9	Christoph Spycher	Switzerland	Eintracht Frankfurt

```
SELECT NATIONAL.PLAYER AS "NIKE
PLAYERS",
FROM NATIONAL, LEAGUE, SPONSORS
WHERE LEAGUE.PLAYER =
NATIONAL.PLAYER
AND SPONSORS.SPONSOR = NIKE;
```

PNL

PNL#	NIKE PLAYERS
PNL1	Philipp Degen
PNL2	Hannu Tihinen
PNL3	Marco Amelia

The Relational Model today

- Today almost all databases are “relational” databases
- “Object-oriented” databases were introduced in the 90s
- They reintroduced a dependency of data and code
- Therefore their influence remained limited

Archiving the Relational Model

- Archiving databases today means mainly archiving relational databases
- In order to preserve all “relations” all tables of a database need to be archived together
- It is not sufficient to convert database tables into a format suitable for long-term preservation. Also the values of the fields of a database must be suitable for long-term preservation:
  - No code
  - No encryption
  - Data types must be suitable for archiving

The SIARD 2.0 Format

- The Swiss Federal Archives has developed a method to store databases as a collection of XML files contained in a compression-less ZIP archive ("ZIP-64" Standard)
- The XML files contain everything that can be stored in a database according to the SQL:1999 standard
- The key to long-term preservation is using open, published, and standardized file formats

The SIARD 2.0 Format

- The SIARD format represents the long term storage format for relational databases

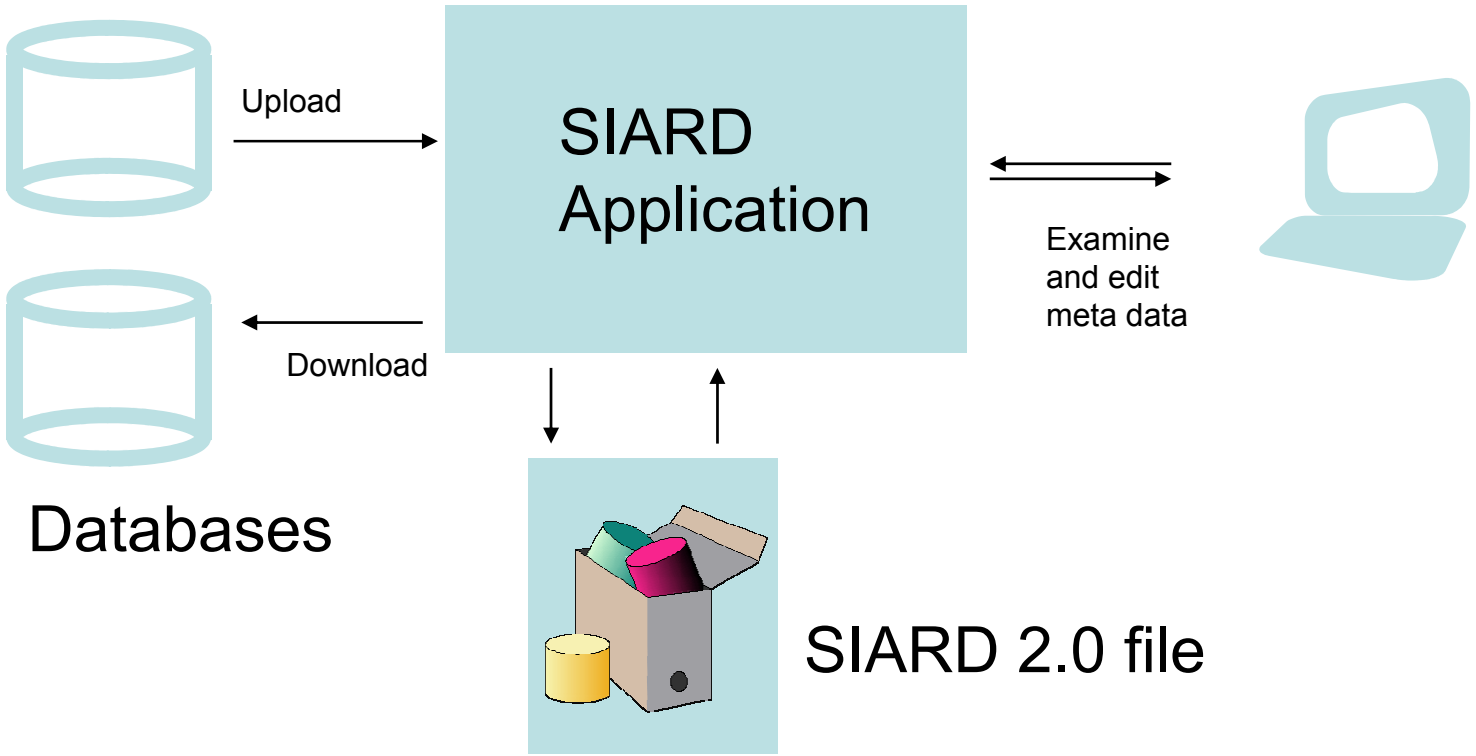
- It is designed independently of package structures such as SIP (Submission Information Package), AIP (Archival Information Package) and DIP (Dissemination Information Package) in the OAIS model

- A database in SIARD format can be archived as part of an archive package which contains additional documents (for understanding the database relevant business records...)

The SIARD 2.0 Format

- The SIARD format stores database content as a SIARD file
- A SIARD file is a ZIP file (ZIP64) containing XML files
- One XML file documents the metadata (based on SQL:1999)
- The other XML files contain the table data
- The SIARD file format is based on open standards:
  SQL:1999, XML, XML Schema, UNICODE, ...

The SIARD 2.0 Format

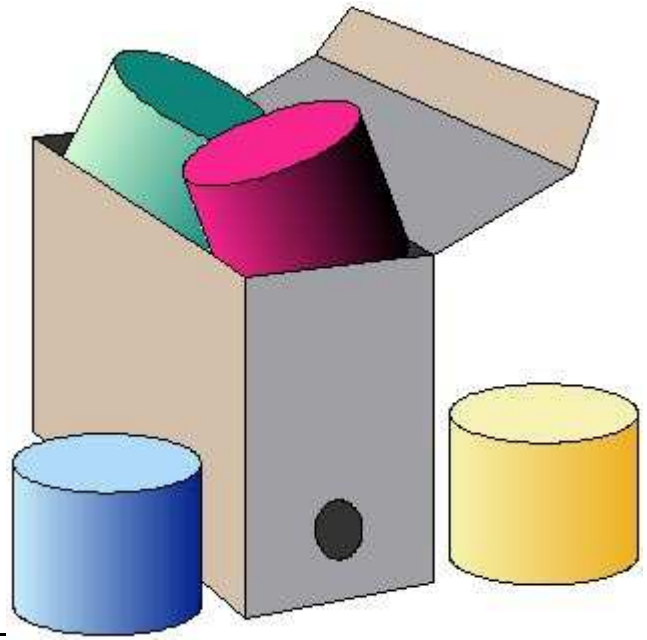


Conclusions

- Most databases to be archived today are relational databases
- Databases are commonly poorly documented by the time they reach the archive
- “Emulation“ of today's database engines for 30 to 50 years appears unrealistic
- Database content must be preserved in files of a standardized, open format: ***SIARD format***
- The SIARD format is based on ISO standards SQL:1999, XML, XML Schema, UNICODE, ...

Part 3
Archiving with SIARD 2.0 – A Presentation of the SIARD Format and Application

Amir Bernstein, PhD

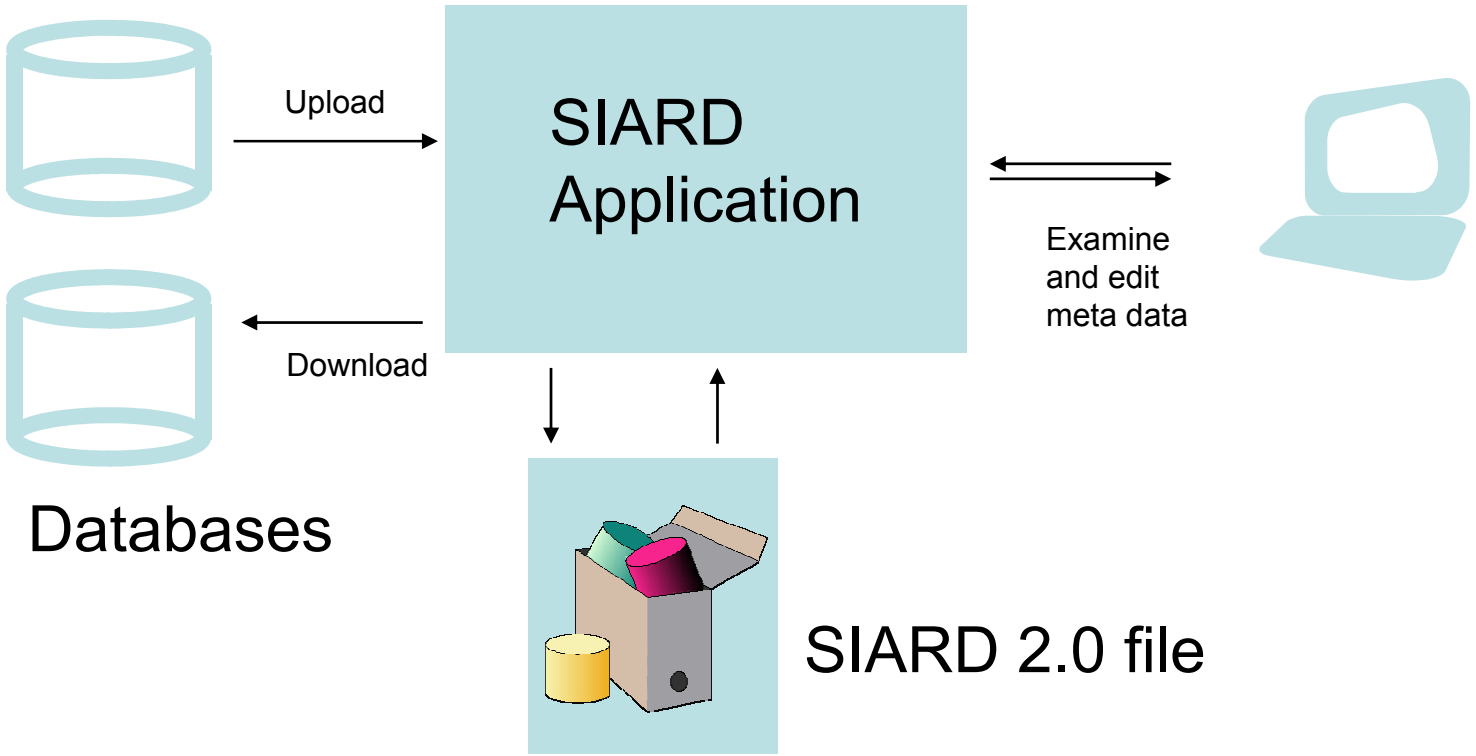


What does SIARD 2.0 do in practical terms?

SIARD is database specific

- **Universal file format** facilitating database management
- SIARD converts databases into an easy-to-handle single file

The SIARD 2.0 Format



Archiving with SIARD

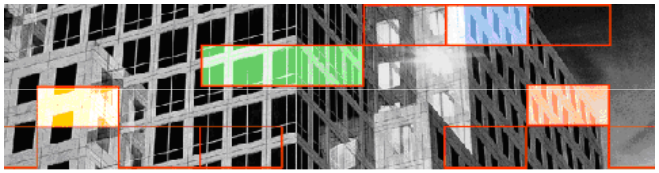
- SIARD converts your and stores your database as a single SIARD file
- SIARD lets you either complete or correct and store the metadata
- The **SFA Package Handler** wraps up your database in an AIP file

Prerequisites I

- **SIARD is platform independent**
  - It operates in a JAVA 1.5 environment
    Windoms, Linux, Mac…
- SIARD can run on a single computer with a common GUI
- **Installation**
  - Use SIARD directly from a USB stick
  - Or simply click & install SIARD

Prerequisites II

SIARD 2.0 is especially conceptualized to support the most widespread databases:

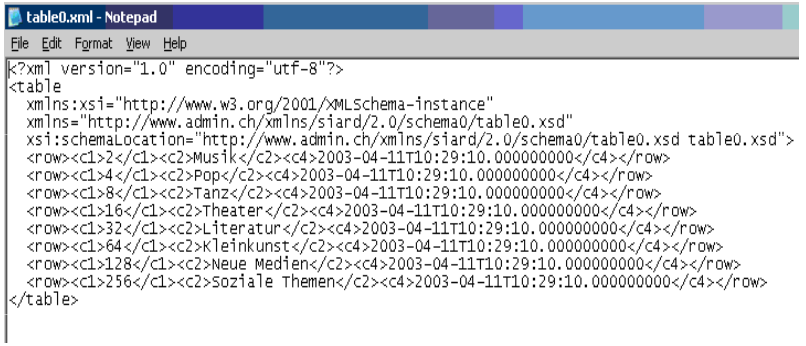


Microsoft SQL Server 2000 Standard Edition

& more…

Primary & Metadata in SIARD

- **Primary data is stored in the folder content**
  - This data will be stored in a XML file format
  - For every table in your database SIARD generates automatically a separate XML file



- **Metadata** is stored one single file: **metadata.xml**

- **Your advantage:**
  - An XML storage will allow you to handle searches within the tables
  - A simple extraction of specific database information according to your needs

SIARD Utilities

SiardEdit

- Edit your metadata
- Create a SIARD-Archive with a new set of metadata
- Match your metadata against those of a different archive
- Update and complete your existing set of metadata
- View and sort your primary data

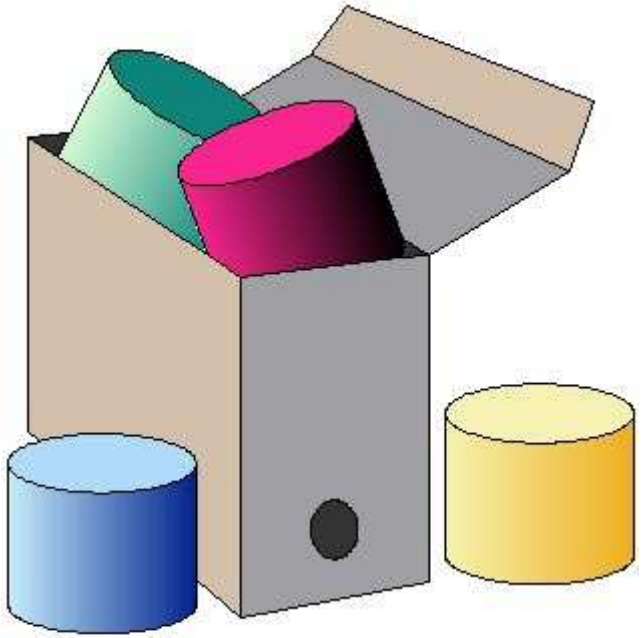
SiardFromDb

- Convert your database into a SIARD-Archive
- Create a full SIARD-Archive (with both metadata and primary data in the SIARD format), or:
- Generate an empty SIARD-Archive (i.e. containing no primary data)

SiardToDb

- Facilitate your research within a given database
- Load your SIARD-Archive into a database instance (with tables, views etc.)
- Comfortably navigate and search within your database

SIARD 2.0 Demo

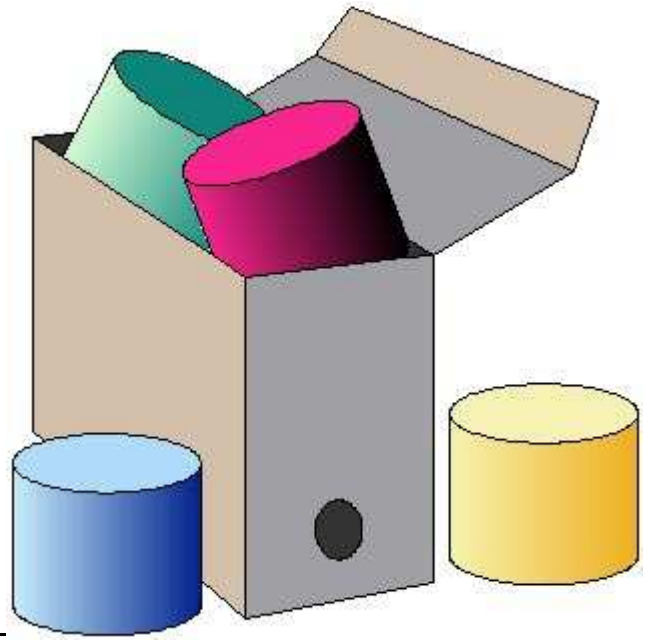


Main strengths of SIARD 2.0

- Offers a universal archiving format
- Is software-invariant
- Conforms with the SQL:1999 and XML standards
- Enables to retain documents' information content

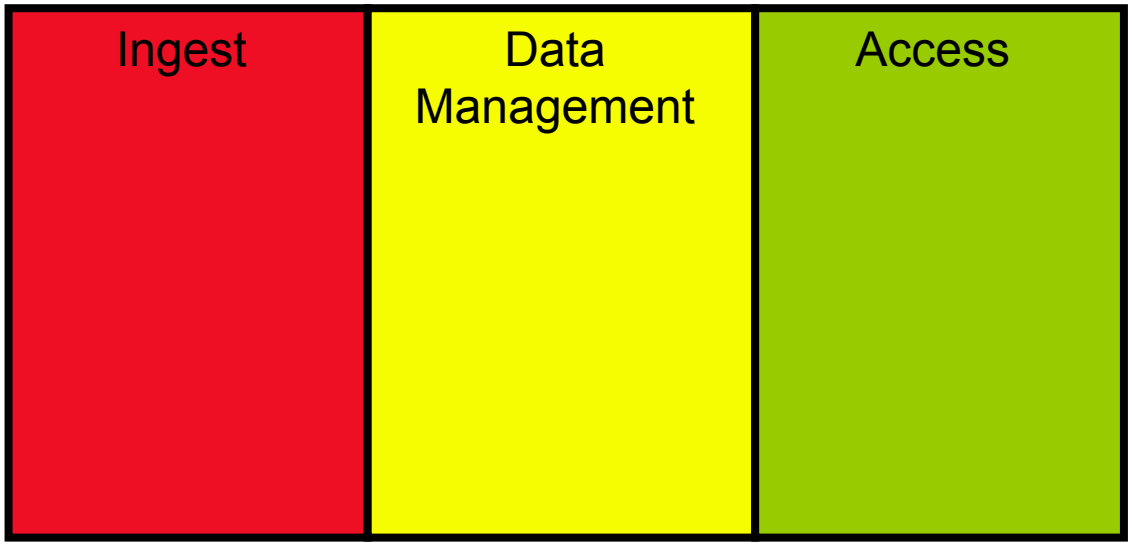
Part 4
Defining a Database Archiving Process

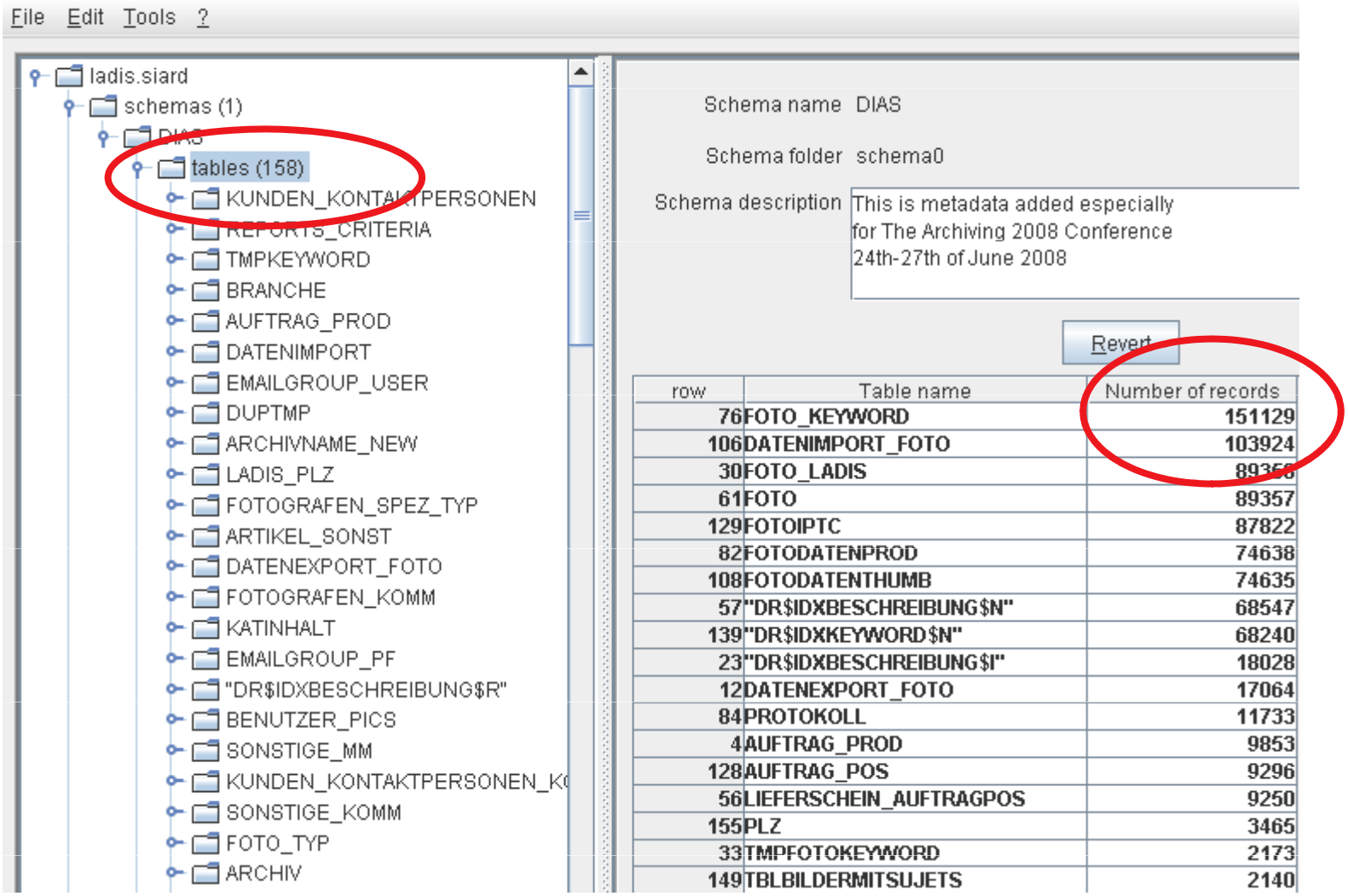
Anne-Louise Joël, MA

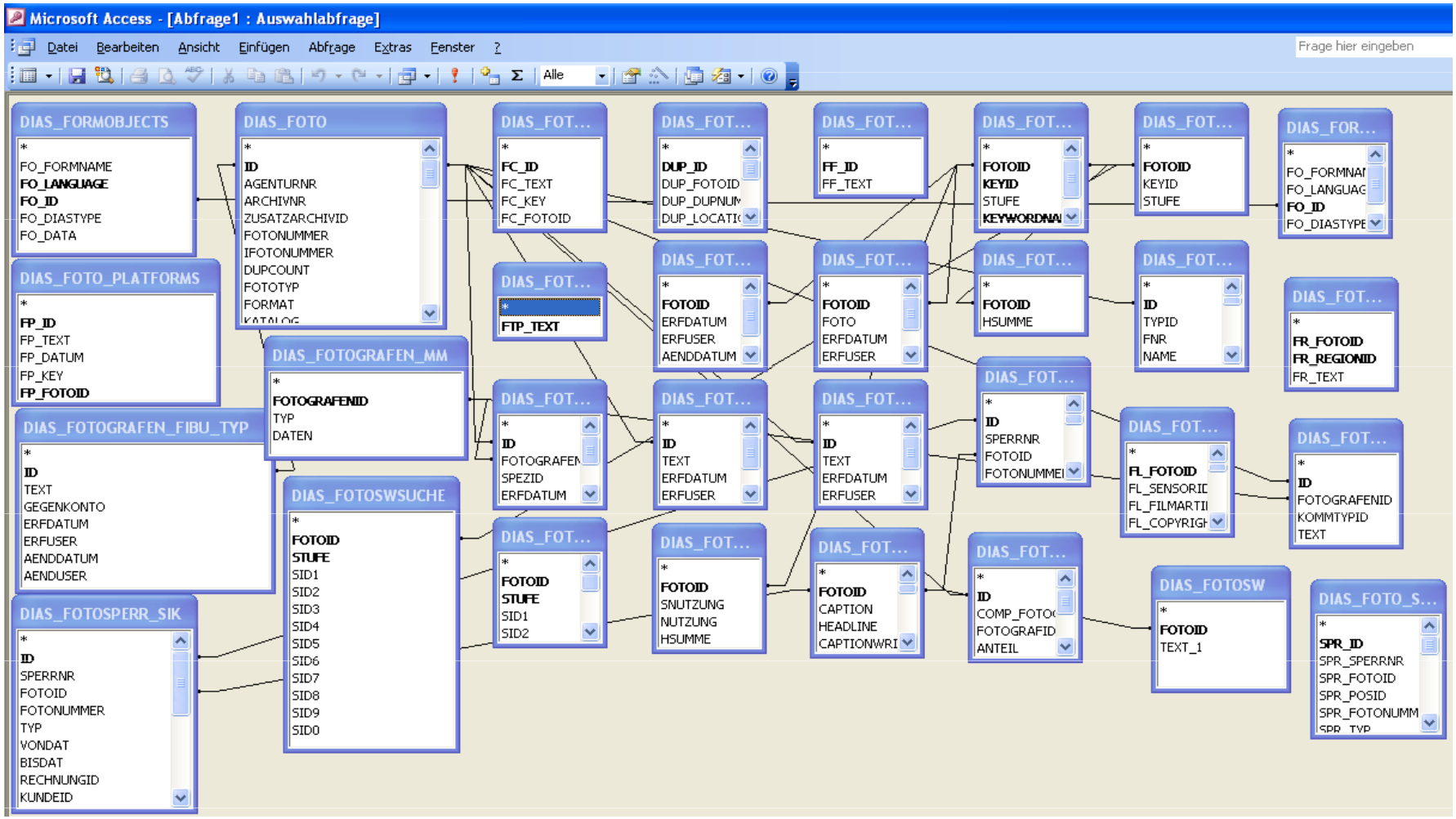


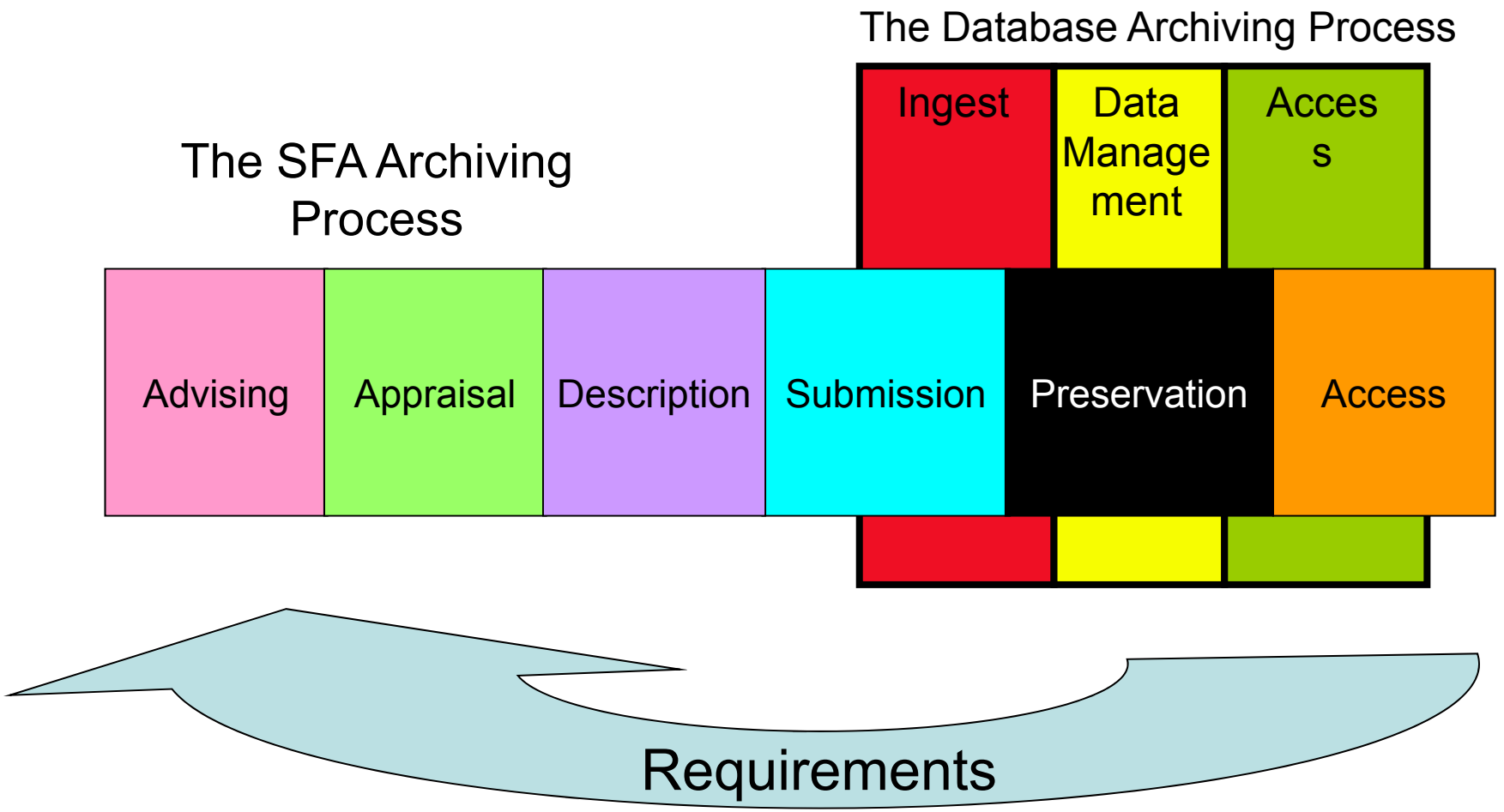
Database Archiving Process

According to the OAIS standard




File Edit Tools ?
ladis.siard
schemas (1)
DIAS
tables (158)
KUNDEN_KONTAKTPERSONEN
REPORTS_CRITERIA
TMPKEYWORD
BRANCHE
AUFTRAG_PROD
DATENIMPORT
EMAILGROUP_USER
DUPTMP
ARCHIVNAME_NEW
LADIS_PLZ
FOTOGRAFEN_SPEZ_TYP
ARTIKEL_SONST
DATENEXPORT_FOTO
FOTOGRAFEN_KOMM
KATINHALT
EMAILGROUP_PF
"DR$IDXBESCHREIBUNG$R"
BENUTZER_PICS
SONSTIGE_MM
KUNDEN_KONTAKTPERSONEN_K
SONSTIGE_KOMM
FOTO_TYP
ARCHIV
Schema name DIAS
Schema folder schema0
Schema description This is metadata added especially for The Archiving 2008 Conference 24th-27th of June 2008
Revert
row | Table name | Number of records
76 | FOTO_KEYWORD | 151129
106 | DATENIMPORT_FOTO | 103924
30 | FOTO_LADIS | 89358
61 | FOTO | 89357
129 | FOTOIPTC | 87822
82 | FOTODATENPROD | 74638
108 | FOTODATENTHUMB | 74635
57 | "DR$IDXBESCHREIBUNG$N" | 68547
139 | "DR$IDXKEYWORD$N" | 68240
23 | "DR$IDXBESCHREIBUNG$I" | 18028
12 | DATENEXPORT_FOTO | 17064
84 | PROTOKOLL | 11733
4 | AUFTRAG_PROD | 9853
128 | AUFTRAG_POS | 9296
56 | LIEFERSCHEIN_AUFTRAGPOS | 9250
155 | PLZ | 3465
33 | TMPFOTOKEYWORD | 2173
149 | TBLBILDERMITSUJETS | 2140


Microsoft Access - [Abfrage1 : Auswahlabfrage]
Datei Bearbeiten Ansicht Einfügen Abfrage Extras Fenster ?
Frage hier eingeben
Alle
DIAS_FORMOBJECTS
FO_FORMNAME
FO_LANGUAGE
FO_ID
FO_DIASTYPE
FO_DATA
DIAS_FOTO
ID
AGENTURNR
ARCHIVNR
ZUSATZARCHIVID
FOTONUMMER
IFOTONUMMER
DUPCOUNT
FOTOTYP
FORMAT
DIAS_FOTO_PLATFORMS
FP_ID
FP_TEXT
FP_DATUM
FP_KEY
FP_FOTOID
DIAS_FOTOGRAFEN_MM
FOTOGRAFENID
TYP
DATEN
DIAS_FOTOGRAFEN_FIBU_TYP
ID
TEXT
GEGENKONTO
ERFDATUM
ERFUSER
AENDDATUM
AENDUSER
DIAS_FOTOSWSUCHE
FOTOID
STUFE
SID1
SID2
SID3
SID4
SID5
SID6
SID7
SID8
SID9
SID0
DIAS_FOTOSPERR_SIK
ID
SPERRNR
FOTOID
FOTONUMMER
TYP
VONDAT
BISDAT
RECHNUNGID
KUNDEID
DIAS_FOT...
FC_ID
FC_TEXT
FC_KEY
FC_FOTOID
FTP_TEXT
DUP_ID
DUP_FOTOID
FOTOID
ERFDATUM
ERFUSER
AENDDATUM
FOTO
FF_ID
FF_TEXT
KEYID
STUFE
HSUMME
TYPID
FNR
NAME
FR_FOTOID
FR_REGIONID
FR_TEXT
SPERRNR
FOTONUMMEI
FL_FOTOID
FL_SENSORID
FL_FILMARTII
FL_COPYRIGH
FOTOGRAFENID
KOMMTYPID
TEXT
SNUTZUNG
NUTZUNG
HSUMME
CAPTION
HEADLINE
CAPTIONWRI
COMP_FOTO
FOTOGRAFID
ANTEIL
DIAS_FOTOSW
FOTOID
TEXT_1
DIAS_FOTO_S...
SPR_ID
SPR_SPERRNR
SPR_FOTOID
SPR_POSID
SPR_FOTONUMM

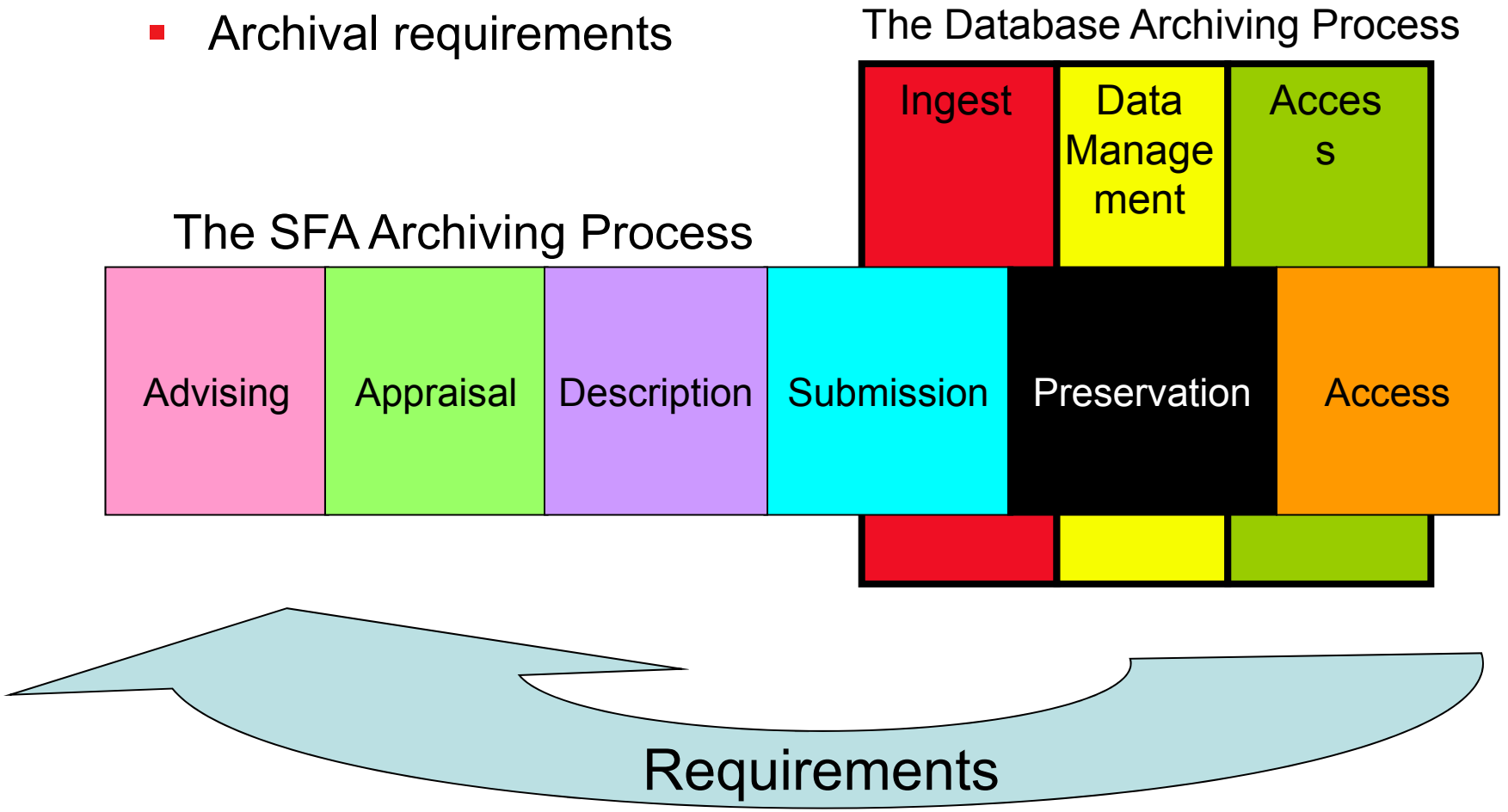

The Database Archiving Process
Ingest
Data Management
Access
The SFA Archiving Process
Advising
Appraisal
Description
Submission
Preservation
Access
Requirements

DAp
Database Archiving Process

- A pilot project at the Swiss Federal Archives
  - on the definition and appliance of the entire archiving process for data from relational databases.

  - The process was defined and tested in detail
  - SFA personnel was educated in database archiving
  - 8 databases were transferred, 13 partially analyzed

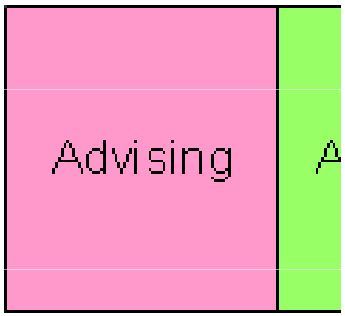
- Technical requirements
- Archival requirements



Advising

- Records management support

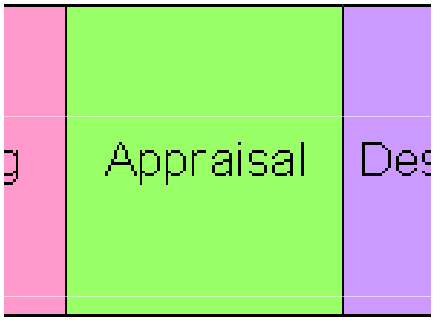
- Before introducing a database application, an agency should consider:
  - Obligation to offer for transfer
  - Concept of disposal
    - Log of changes
    - Historization
  - Technical documentation of the database
    - Visual Data model
    - Data dictionary



Appraisal

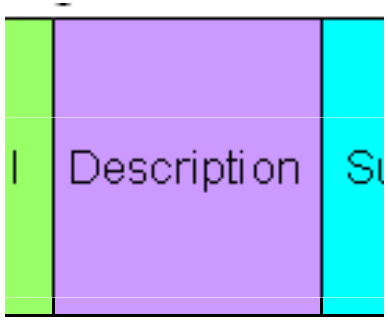
- Archival Appraisal
  - The evidential value
  - The information value

- Technical Appraisal
  - Is it a relational database?
  - The amount of tables
  - The amount of records



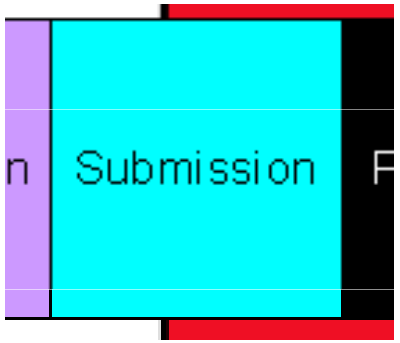
Description

- Technical metadata
  - Amount and names tables
  - Type of data, etc.
- Archival metadata
  - Context
  - Documentation
    - Educational material
    - Screen shots
    - Data dictionary
- Administrative metadata
  - information on the delivered object



Submission

- Archive database at the agency for interim storage?
- Snapshot or deleted data?
- Original or simplified version of database?
- Depending on
  - Legal prerequisites
  - Technical feasibility

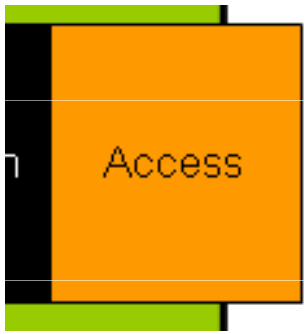


Preservation

- Technical preservation with SIARD 2.0
- Migration strategies
  - ISO SIARD

Access

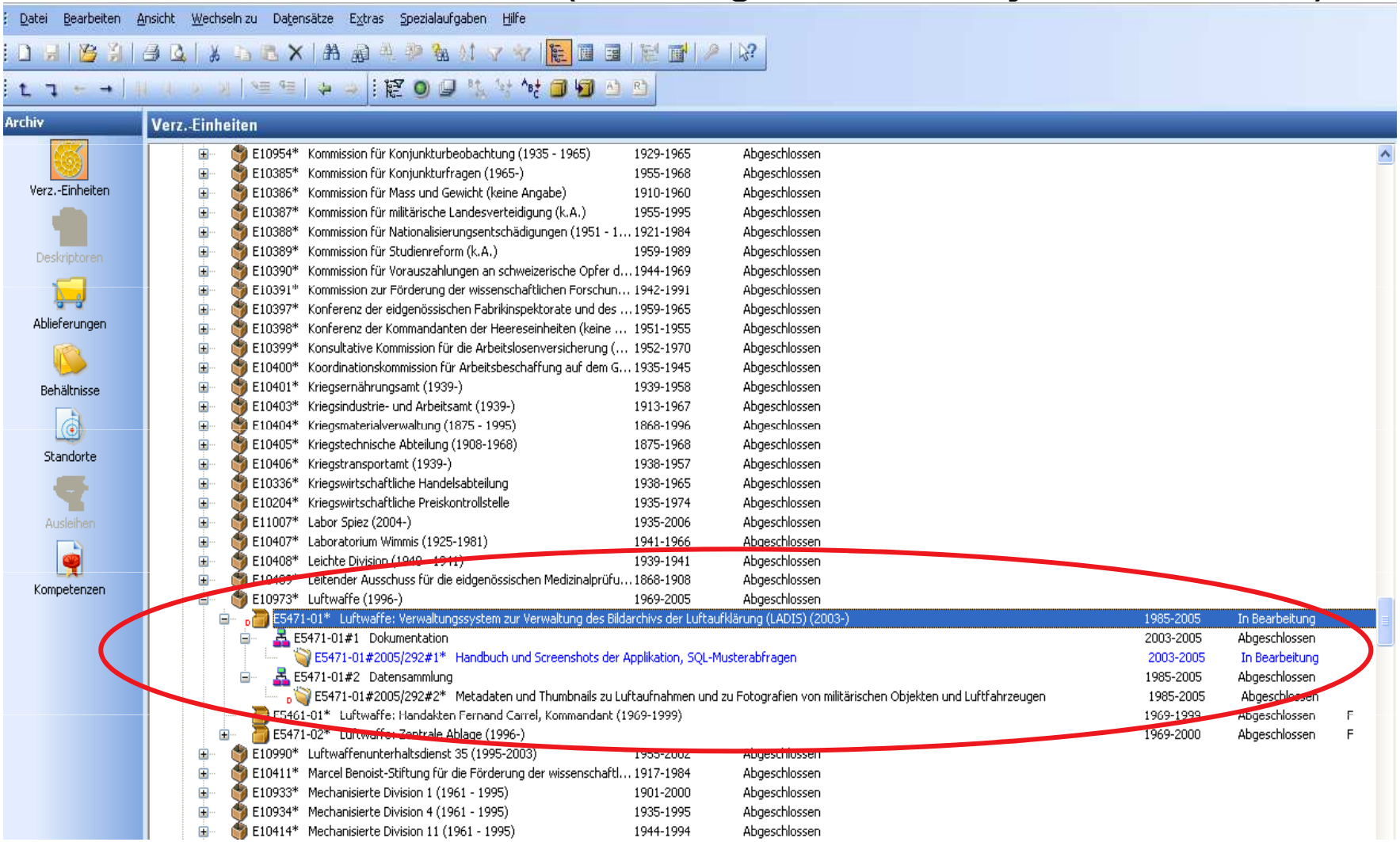
- Accessing a database at the Swiss Federal Archives



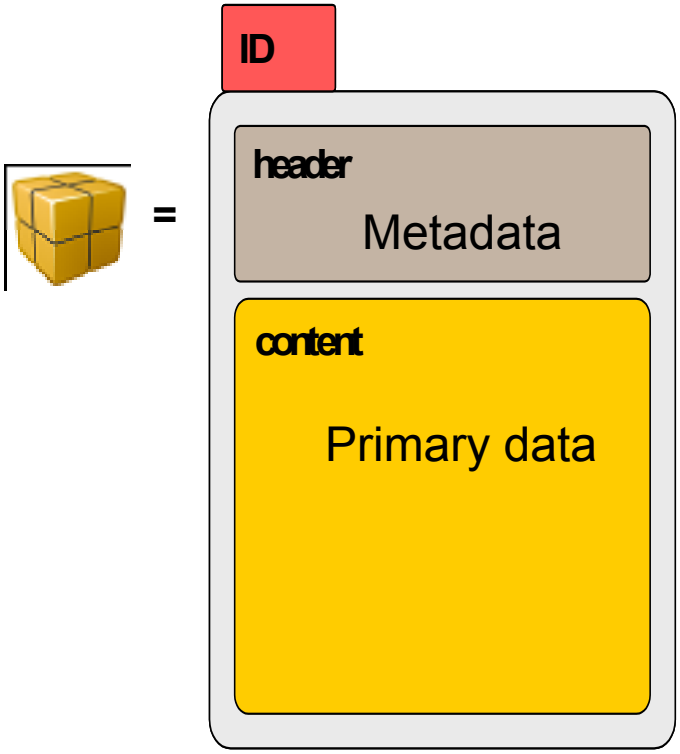
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Schweizerisches Bundesarchiv BAR
Ressort Innovation und Erhaltung

LADIS in AIS (Archiving Information System at the SFA)



SIARD within a SFA Information Package



=

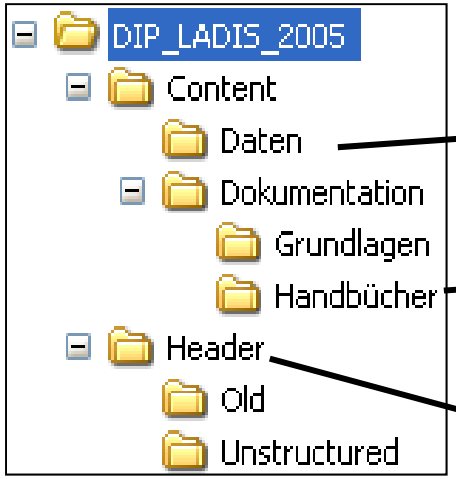


Table name DATENEXPORT_FOTO

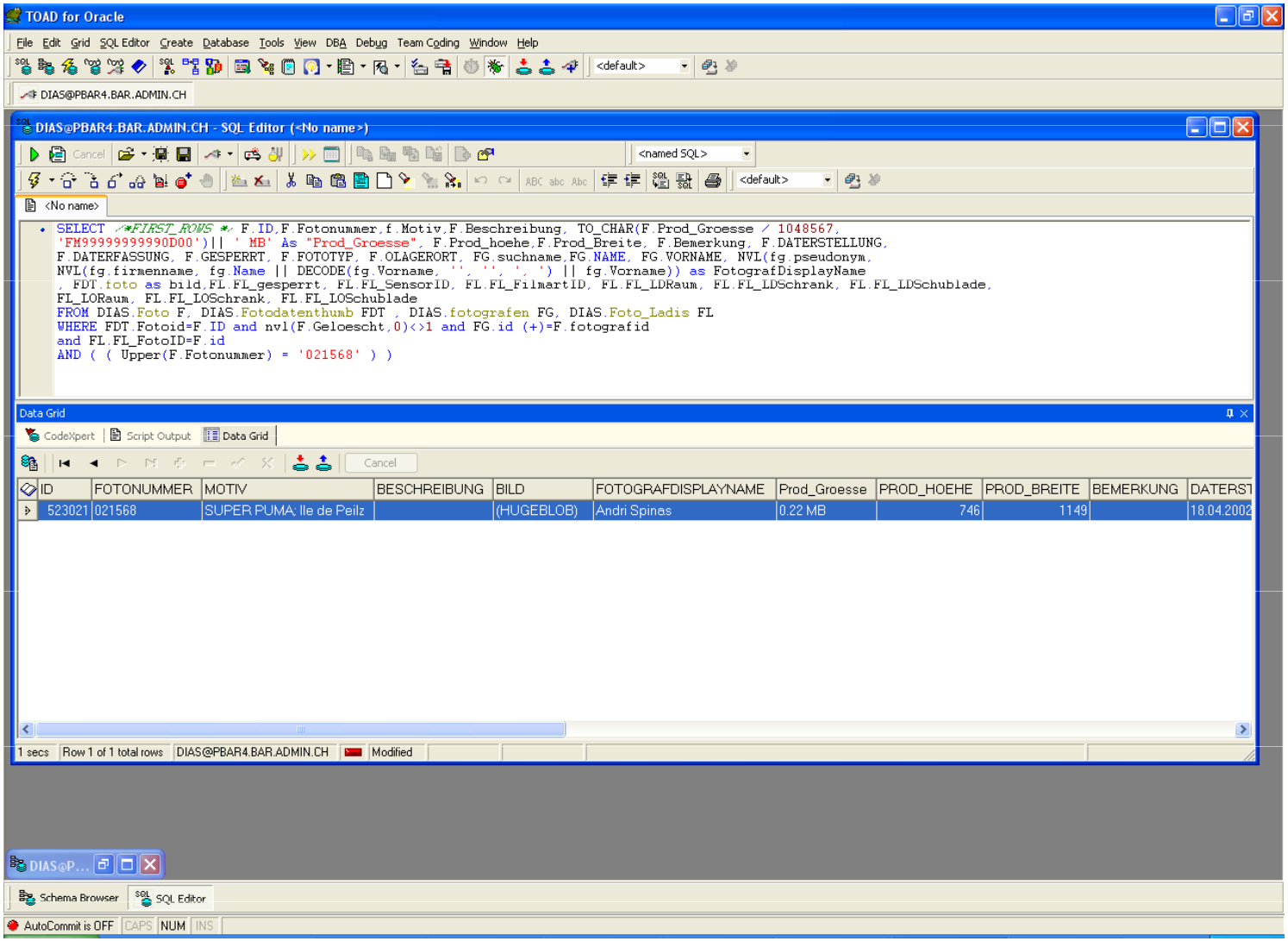
Records 0 - 15000 (2500)

row	DEF_FOTOID	DEF_DATENEXPORTID	DEF
00000	1000003.0	1.0	
02500	580405.0	279.0	
05000	523917.0	400.0	
07500	571520.0	599.0	
10000	527807.0	719.0	
12500	542817.0	850.0	
15000	1014097.0	982.0	

ladis.siard

Handbuch LADIS Version 2.pdf
handbuch LADIS.pdf

metadata.xml


TOAD for Oracle
File Edit Grid SQL Editor Create Database Tools View DBA Debug Team Coding Window Help
DIAS@PBAR4.BAR.ADMIN.CH
DIAS@PBAR4.BAR.ADMIN.CH - SQL Editor (<No name>)
<No name>
SELECT /*FIRST_ROWS */ F.ID,F.Fotonummer,f.Motiv,F.Beschreibung, TO_CHAR(F.Prod_Groesse / 1048567,
'FM99999999990D00')|| ' MB' As "Prod_Groesse", F.Prod_hoehe,F.Prod_Breite, F.Bemerkung, F.DATERSTELLUNG,
F.DATERFASSUNG, F.GESPERRT, F.FOTOTYP, F.OLAGERORT, FG.suchname,FG.NAME, FG.VORNAME, NVL(fg.pseudonym,
NVL(fg.firmenname, fg.Name || DECODE(fg.Vorname, '', '', ', ') || fg.Vorname)) as FotografDisplayName
, FDT.toto as bild,FL.FL_gesperrt, FL.FL_SensorID, FL.FL_FilmartID, FL.FL_LDRaum, FL.FL_LDSchrank, FL.FL_LDSchublade,
FL_LORaum, FL.FL_LOSchrank, FL.FL_LOSchublade
FROM DIAS.Foto F, DIAS.Fotodatenthumb FDT , DIAS.fotografen FG, DIAS.Foto_Ladis FL
WHERE FDT.Fotoid=F.ID and nvl(F.Geloescht,0)<>1 and FG.id (+)=F.fotografid
and FL.FL_FotoID=F.id
AND ((Upper(F.Fotonummer) = '021568'))
Data Grid
CodeXpert Script Output Data Grid
Cancel
ID FOTONUMMER MOTIV BESCHREIBUNG BILD FOTOGRAFDISPLAYNAME Prod_Groesse PROD_HOEHE PROD_BREITE BEMERKUNG DATERST
523021 021568 SUPER PUMA: Ile de Peilz (HUGEBLOB) Andri Spinas 0.22 MB 746 1149 18.04.2002
1 secs Row 1 of 1 total rows DIAS@PBAR4.BAR.ADMIN.CH Modified
Schema Browser SQL Editor
AutoCommit is OFF CAPS NUM INS

```
SELECT /*FIRST_ROWS */ F.ID,F.Fotonummer,f.Motiv,F.Beschreibung, TO_CHAR(F.Prod_Groesse / 1048567,
'FM99999999990D00')|| ' MB'  As "Prod_Groesse", F.Prod_hoehe,F.Prod_Breite, F.Bemerkung, F.DATERSTELLUNG,
F.DATERFASSUNG, F.GESPERRT, F.FOTOTYP, F.OLAGERORT, FG.suchname,FG.NAME, FG.VORNAME, NVL(fg.pseudonym,
NVL(fg.firmenname, fg.Name || DECODE(fg.Vorname, '', '', ', ') || fg.Vorname)) as FotografDisplayName
, FDT.foto as bild,FL.FL_gesperrt, FL.FL_SensorID, FL.FL_FilmartID, FL.FL_LDRaum, FL.FL_LDSchrank, FL.FL_LDSchublade,
FL_LORaum, FL.FL_LOSchrank, FL.FL_LOSchublade
FROM  ladis.Foto F,ladis.Fotodatenthumb FDT ,ladis.fotografen FG, ladis.Foto_Ladis FL
  WHERE FDT.Fotoid=F.ID and nvl(F.Geloescht,0)<>1 and FG.id (+)=F.fotografid
  and FL.FL_FotoID=F.id
  AND ( ( Upper(F.Fotonummer) = '021568' ) )
```

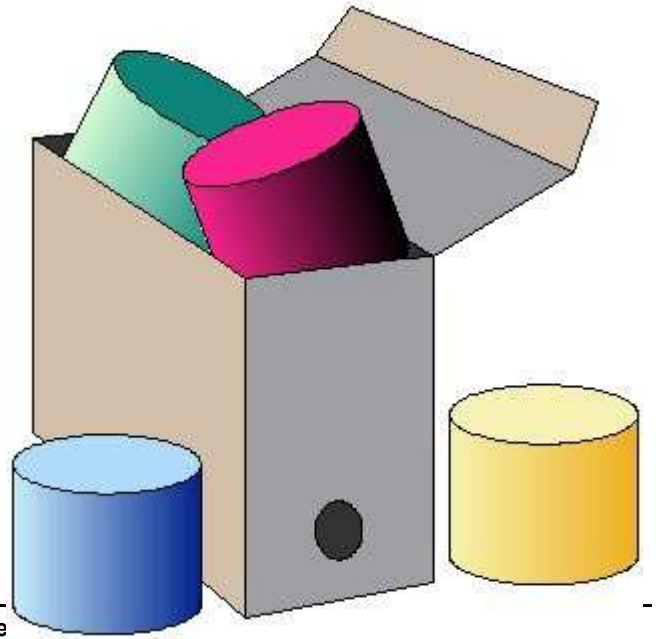
The Database Archiving Process

- Objectives:
  - To secure the access to the primary data
  - To set up and represent the primary data in a user friendly and well-structured environment
  - Possibility to capture and update metadata
  - Possibility to conduct research within the primary data

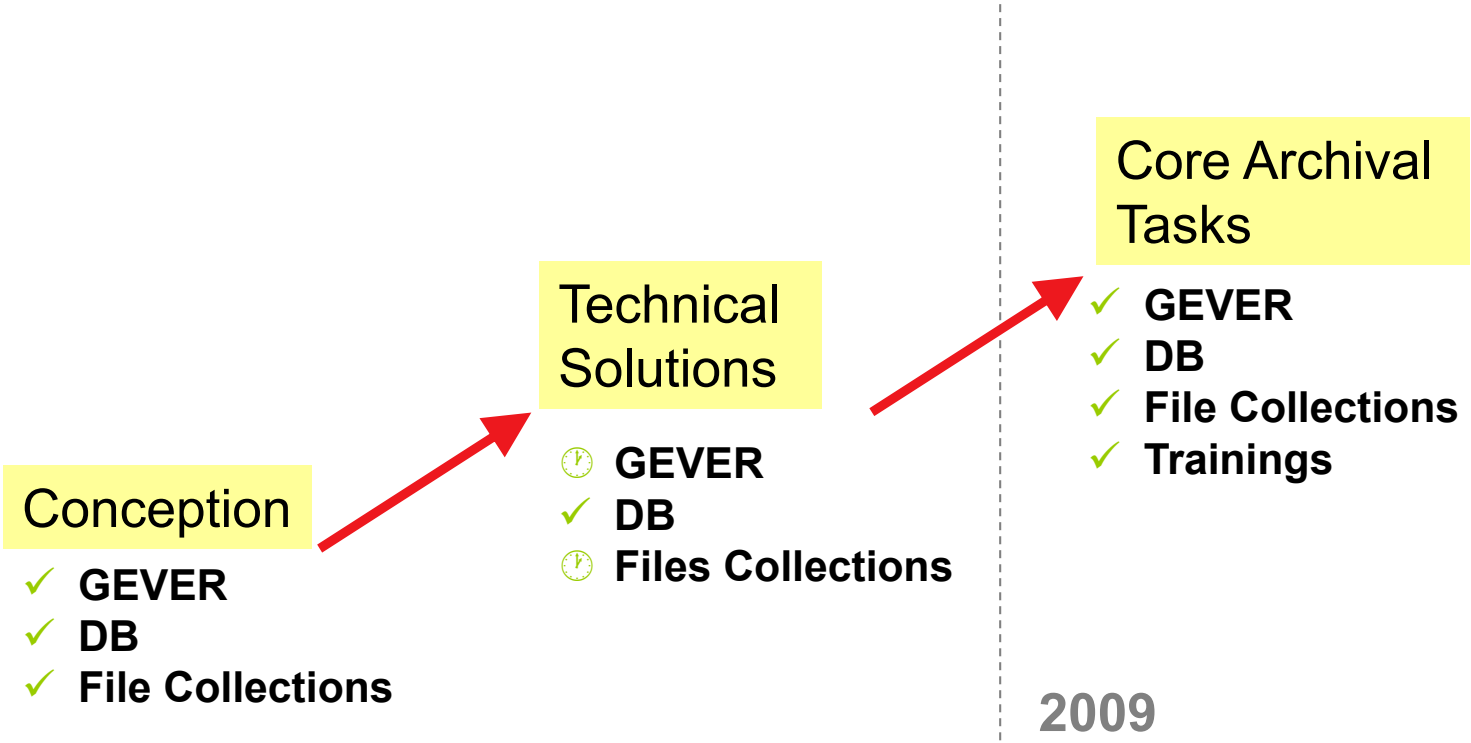
- Advantages:
  - Agencies: Standard process, reliability and efficiency during the submission phase
  - The Swiss Federal Archives: Ensuring readability of data types, encoding and context
  - Users: Possibility of conducting research by different criteria

Wrap Up

- Outlook

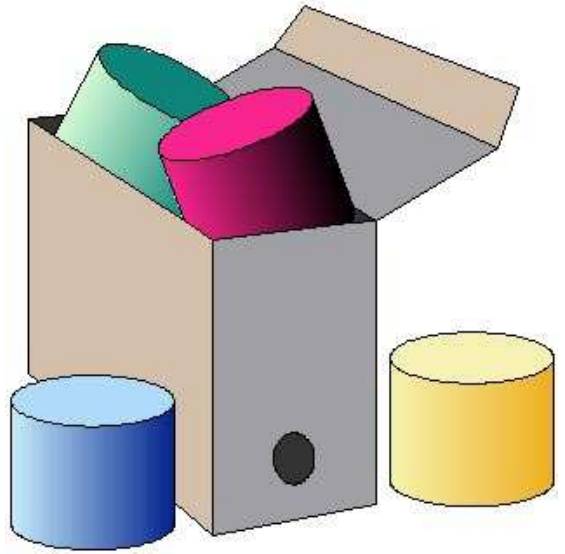


Standpoint ARELDA



SIARD 2.0 – Perspectives

- SIARD became in May 2008 the official format of the PLANETS project
- Further development of SIARD software
- Take-up tests until September 2008
- Dissemination of SIARD 2.0 among the Swiss federal offices and agencies by the end of 2008



Contact us:

bundesarchiv@bar.admin.ch

Swiss Federal Archives
Archiv St. 24
CH-3003 Berne
SWITZERLAND

Phone: + 41 31 322 89 89

Fax: + 41 31 322 78 23