

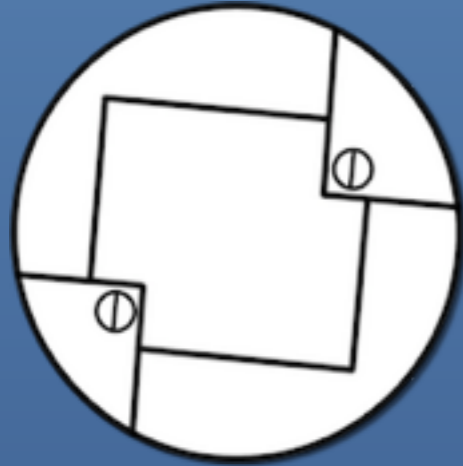
Database Preservation Toolkit

State-of-the-art and future plans
in the context of the E-ARK project

Is a Data Warehouse a Data Archive and Why Does It Matter?

**Hélder Silva hsilva@keep.pt
KEEP SOLUTIONS**

February 19-20, 2015



RODA
REPOSITORY OF AUTHENTIC DIGITAL OBJECTS



Open source
Vendor freedom, no licensing costs



Ready to use
By archivists, producers and consumers



Authenticity
Preservation metadata, ISO 16363 compliant

Database Preservation Toolkit

Developed within RODA project (2006)

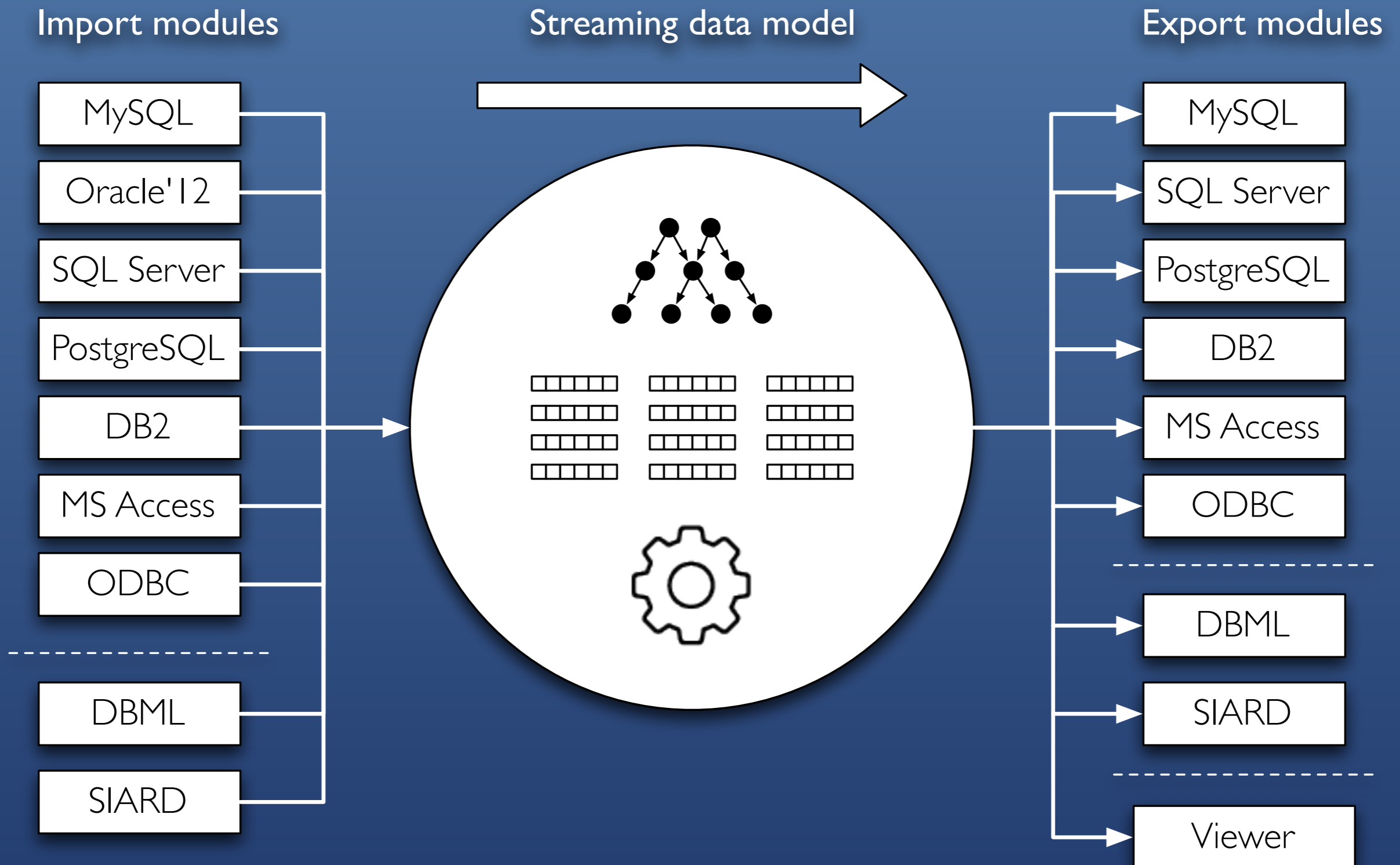
Now stand-alone open-source project (2014)

<http://keeps.github.io/db-preservation-toolkit/>

Imports and exports between DBMS and DB formats

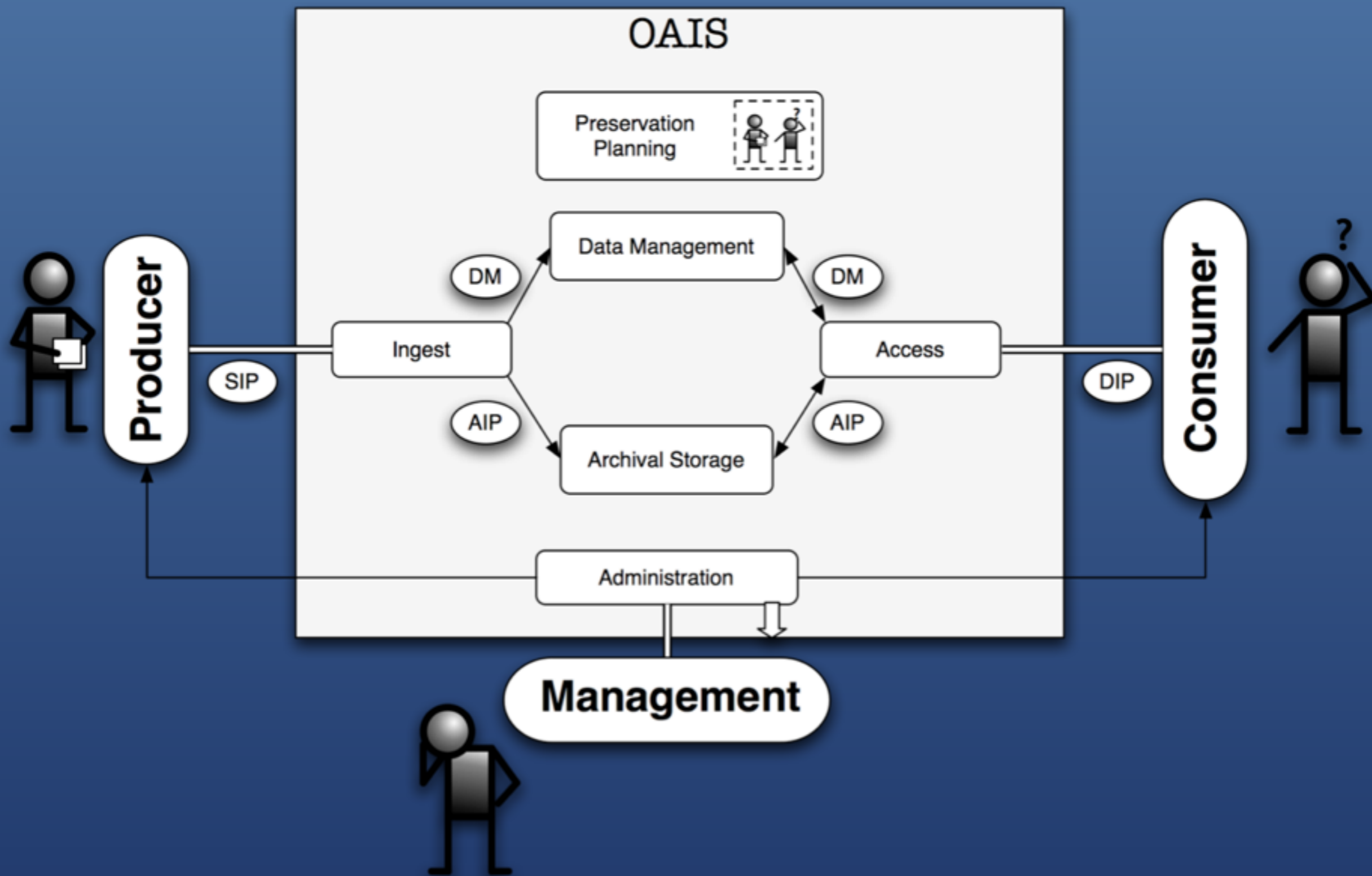
Supports preservation formats: DBML, SIARD

db-preservation-toolkit architecture

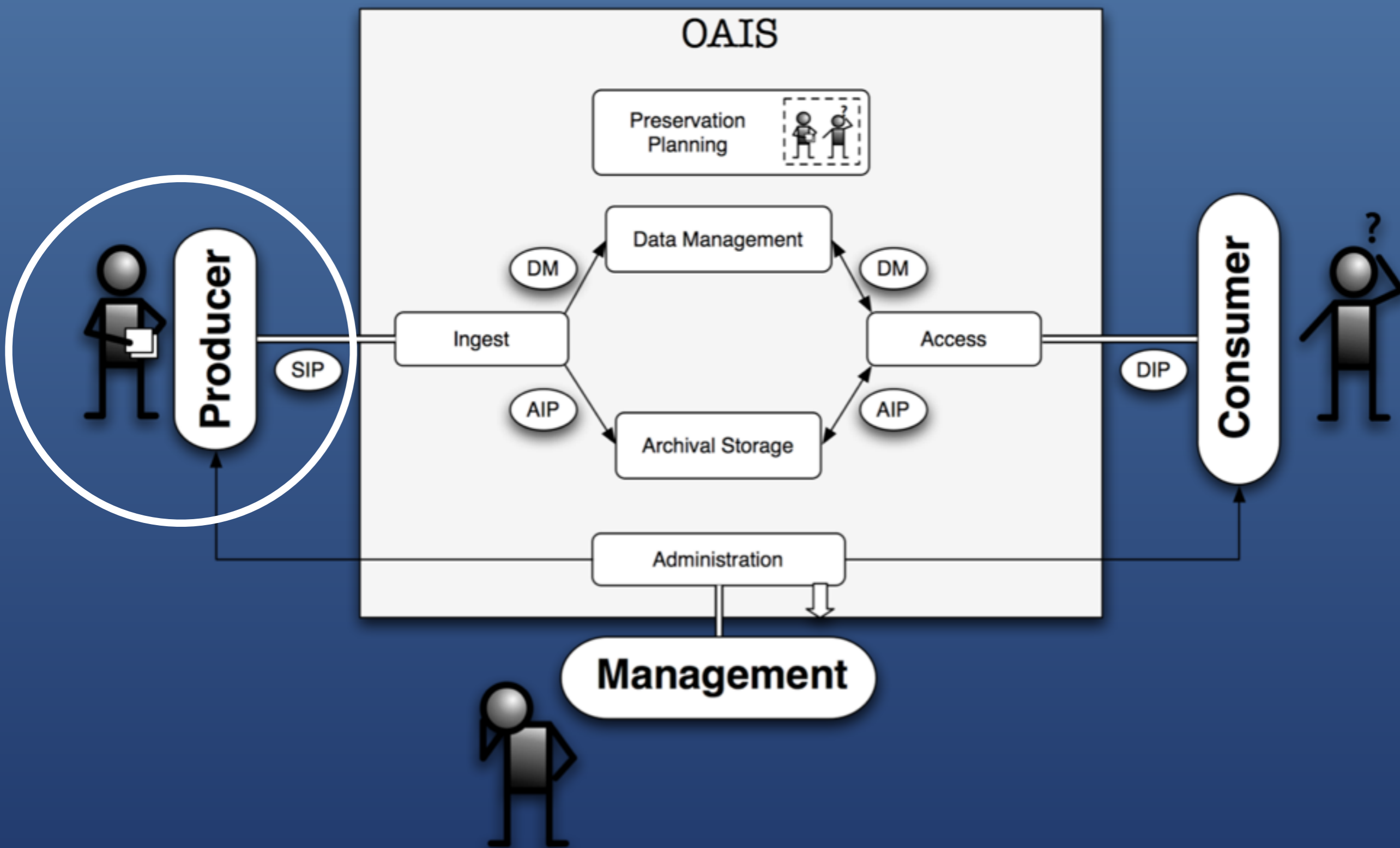


In the Archive

In the Archive

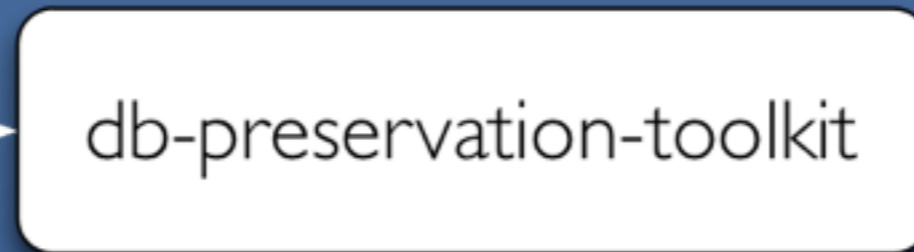
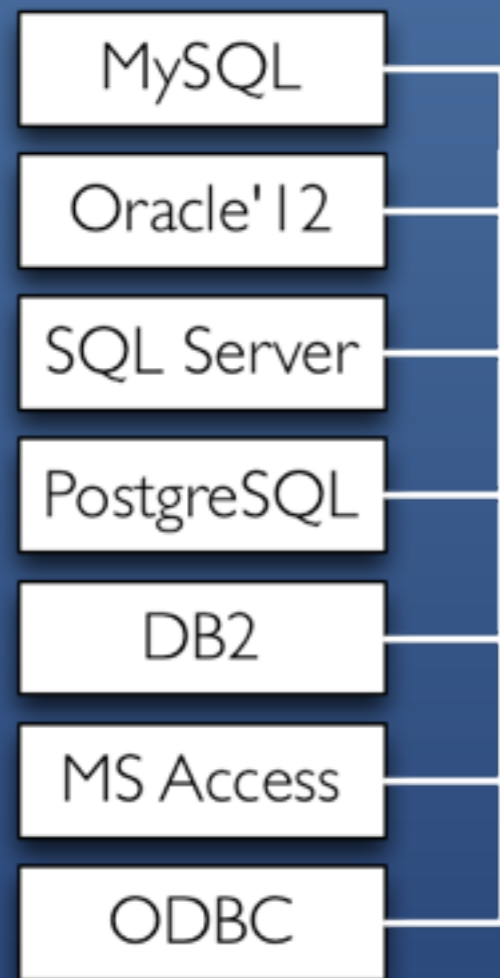


In the Archive



Pre-ingest / Production / SIP creation

Origin formats

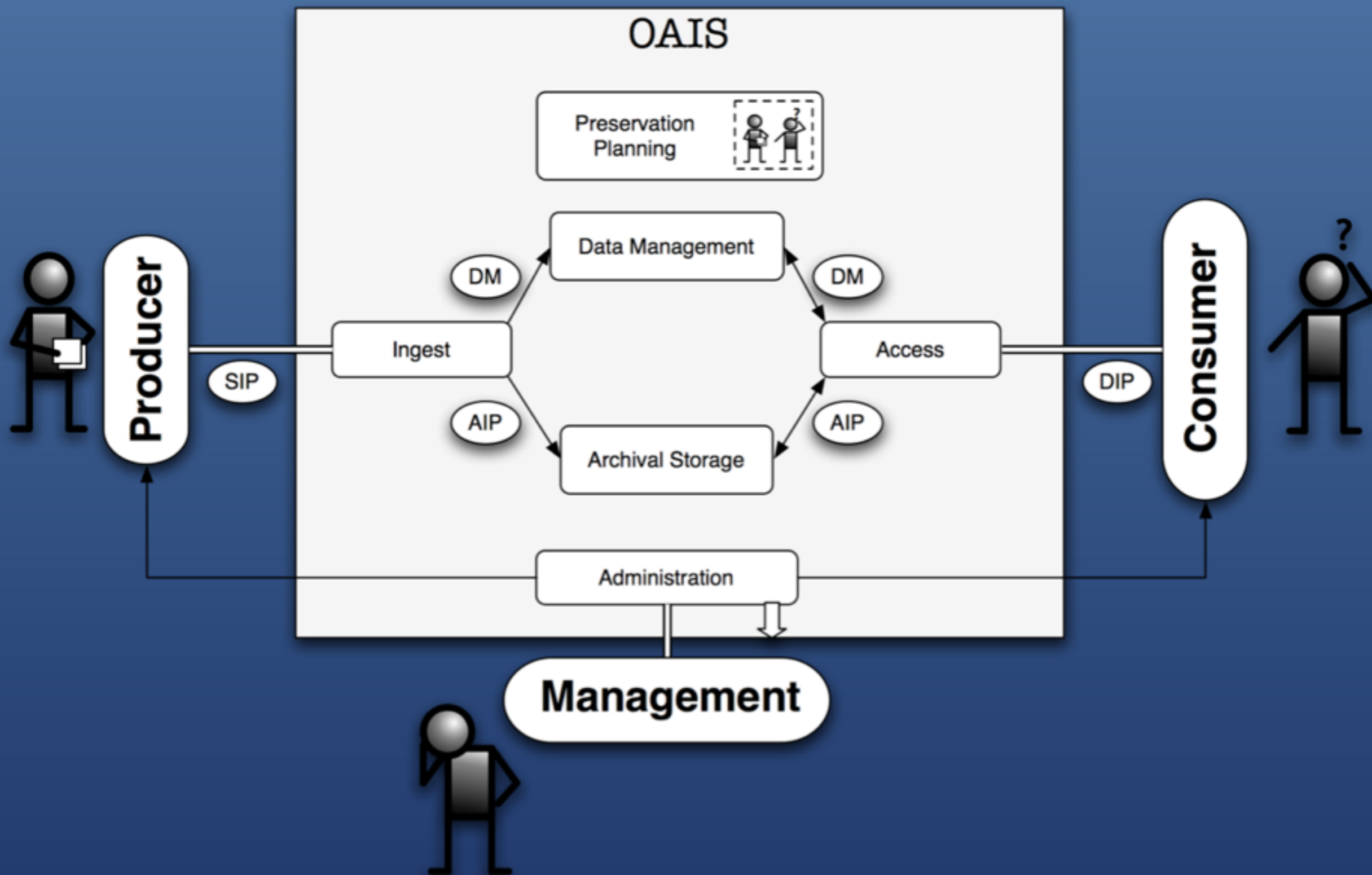


Preservation formats

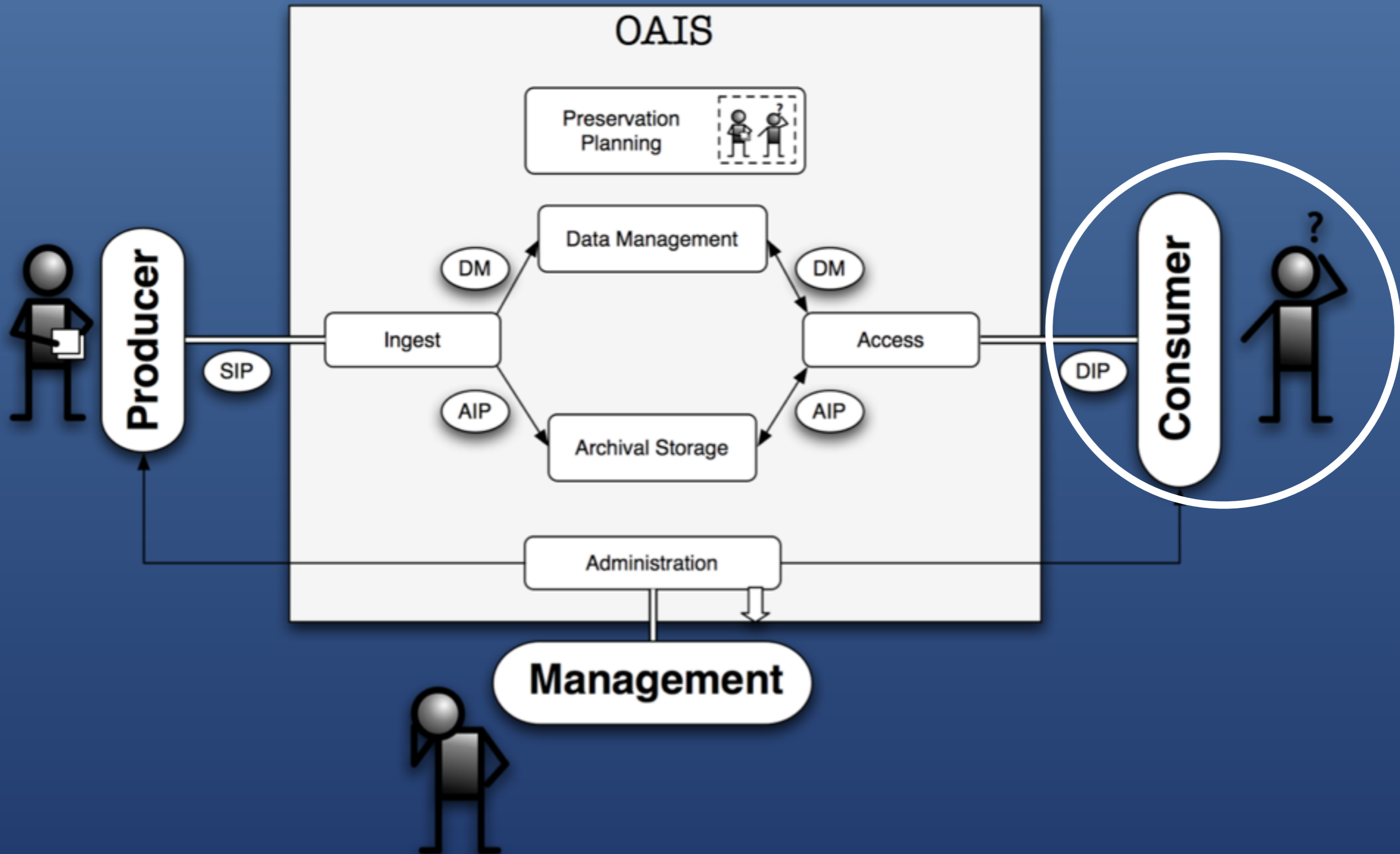


Back to the Archive

Back to the Archive



Back to the Archive



Access

Preservation formats

DBML

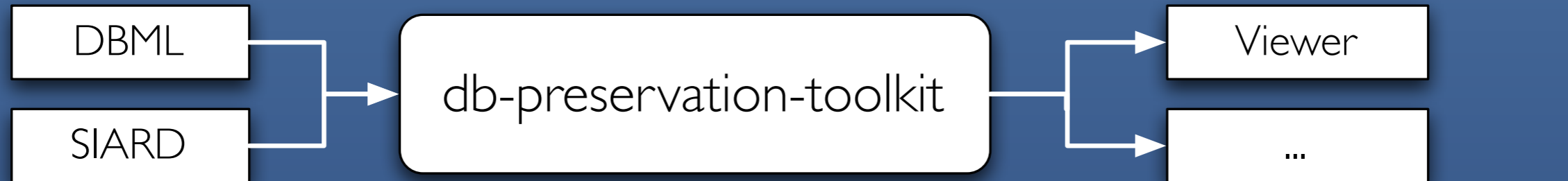
SIARD

db-preservation-toolkit

Access format / viewer

Viewer

...



Preservation formats: DBML

First preservation format

Plain XML

- Associated XML Schema

- Software/hardware independent

- Easy to process

- Descriptive

- Extensible

Stores the data and database structure in a single XML file

Never meant to be a standard but no “standard” was available at the time (2006)

Preservation formats: SIARD

Developed by the Swiss Federal Archives

SIARD

Suite: tools to create/view

Format: open file format specification

Standards based (XML, SQL:1999, Unicode)

Stores everything in an uncompressed ZIP file

Swiss standard since 2013 (eCH-0165)

Preservation formats: SIARD

Added in version 2.0

Both input & output module

Generated output was validated using KOST-Val

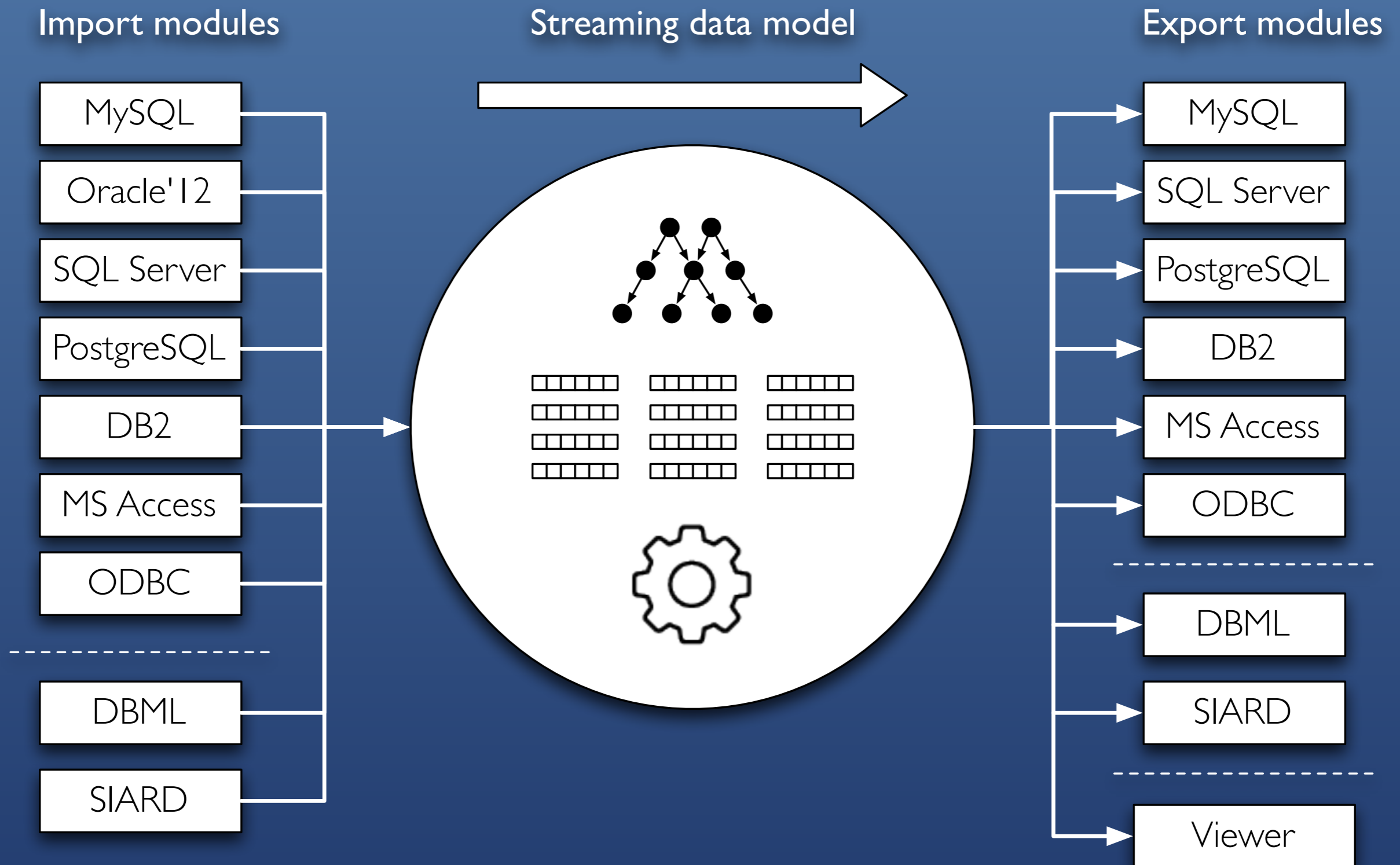
<https://github.com/KOST-CECO/KOST-Val>

Validating SIARD

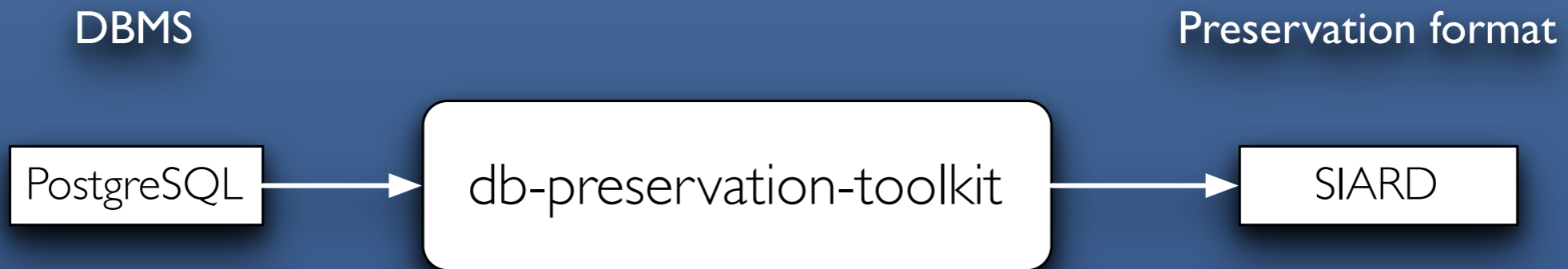
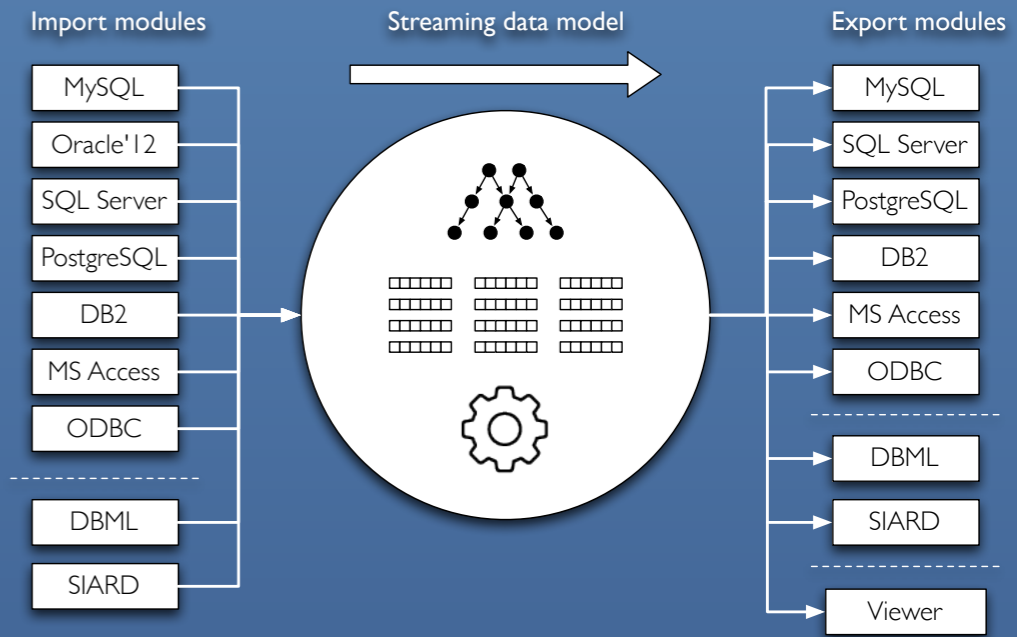
Validating SIARD



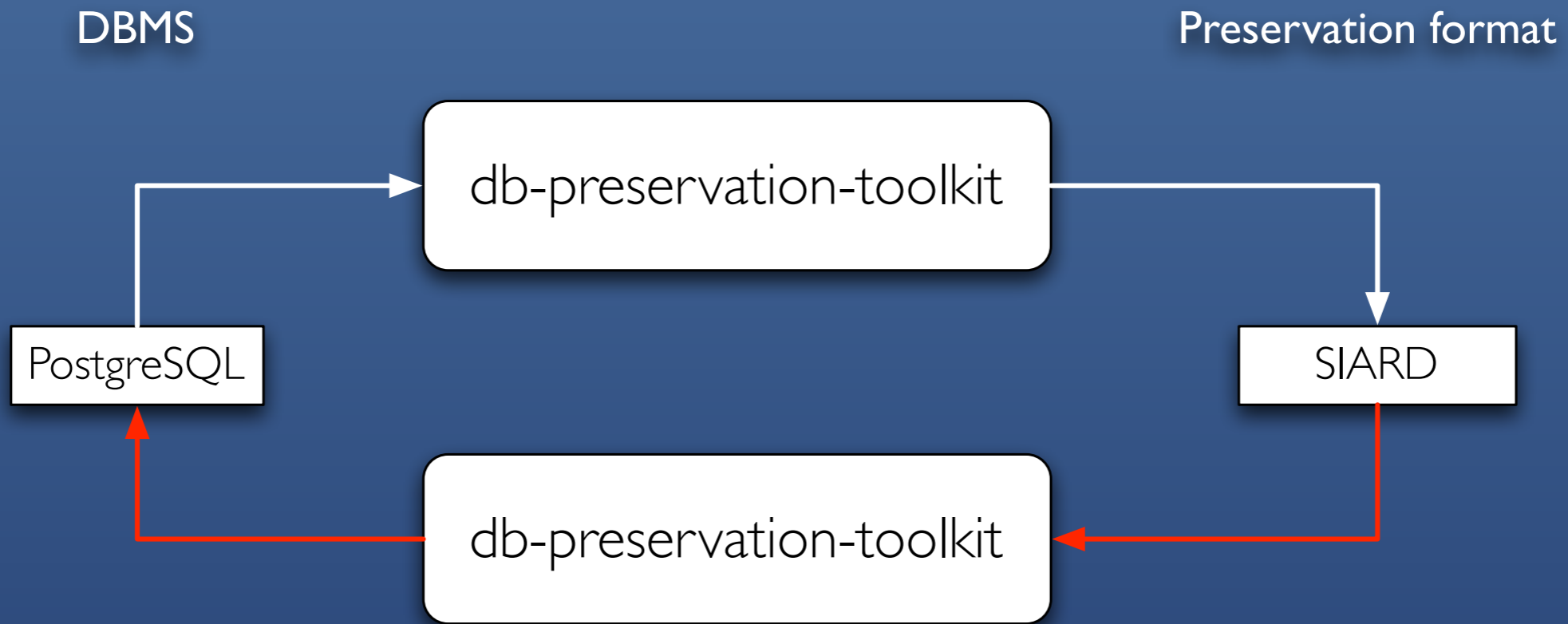
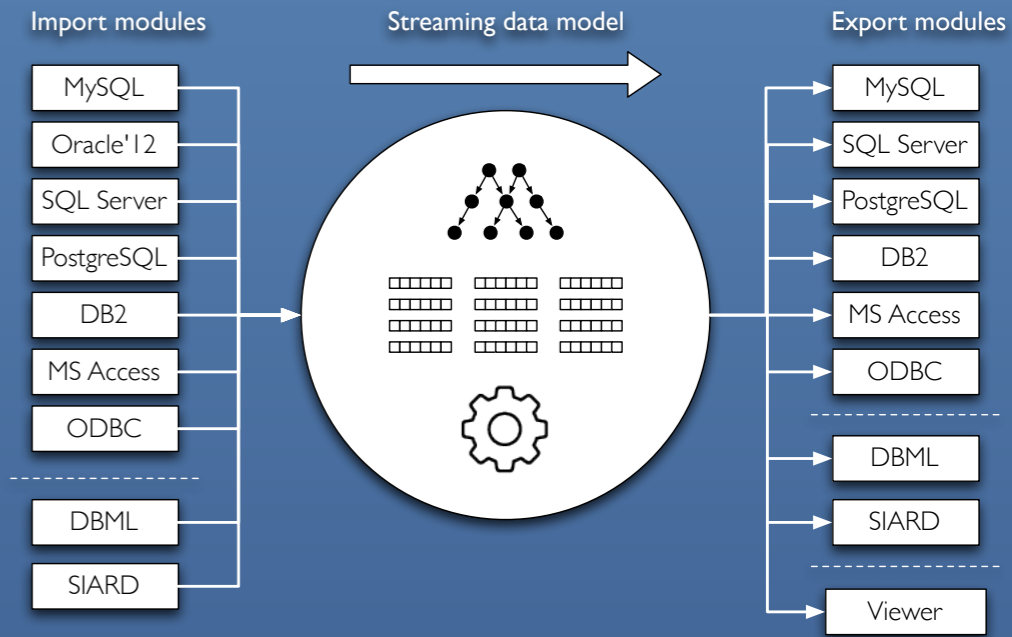
Validating SIARD



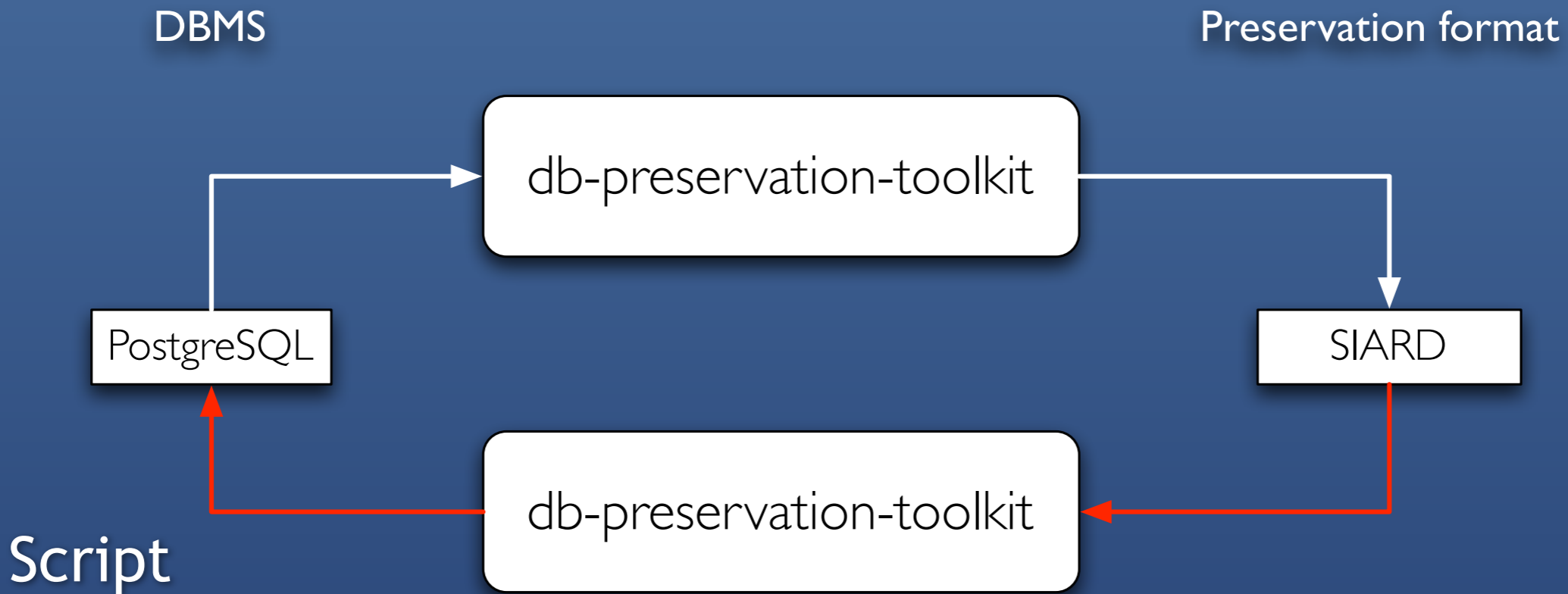
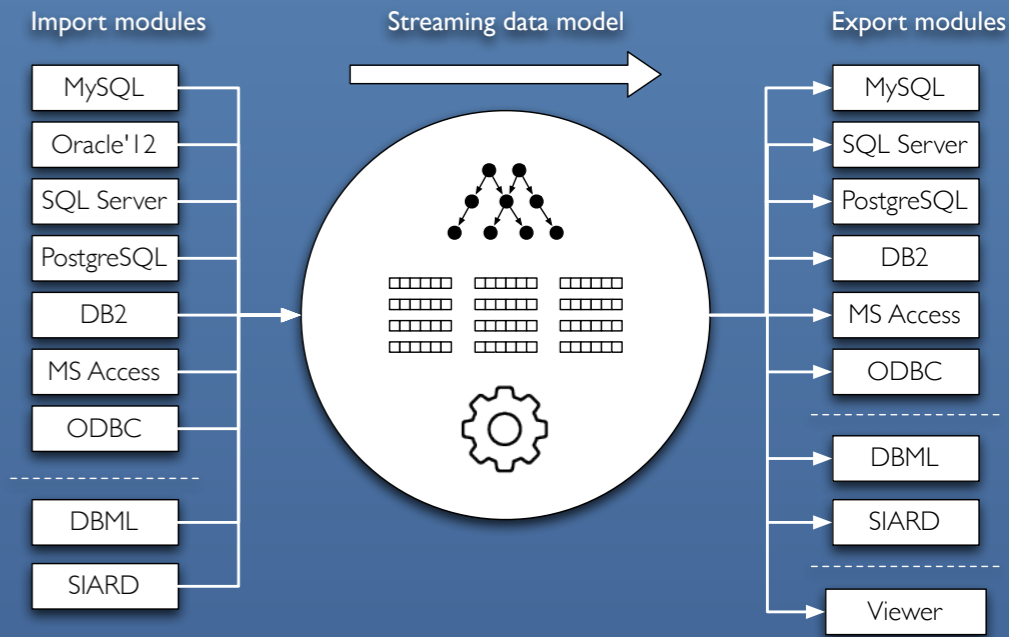
Validating SIARD



Validating SIARD



Validating SIARD



Database initialization (create database, insert data with all datatypes available)

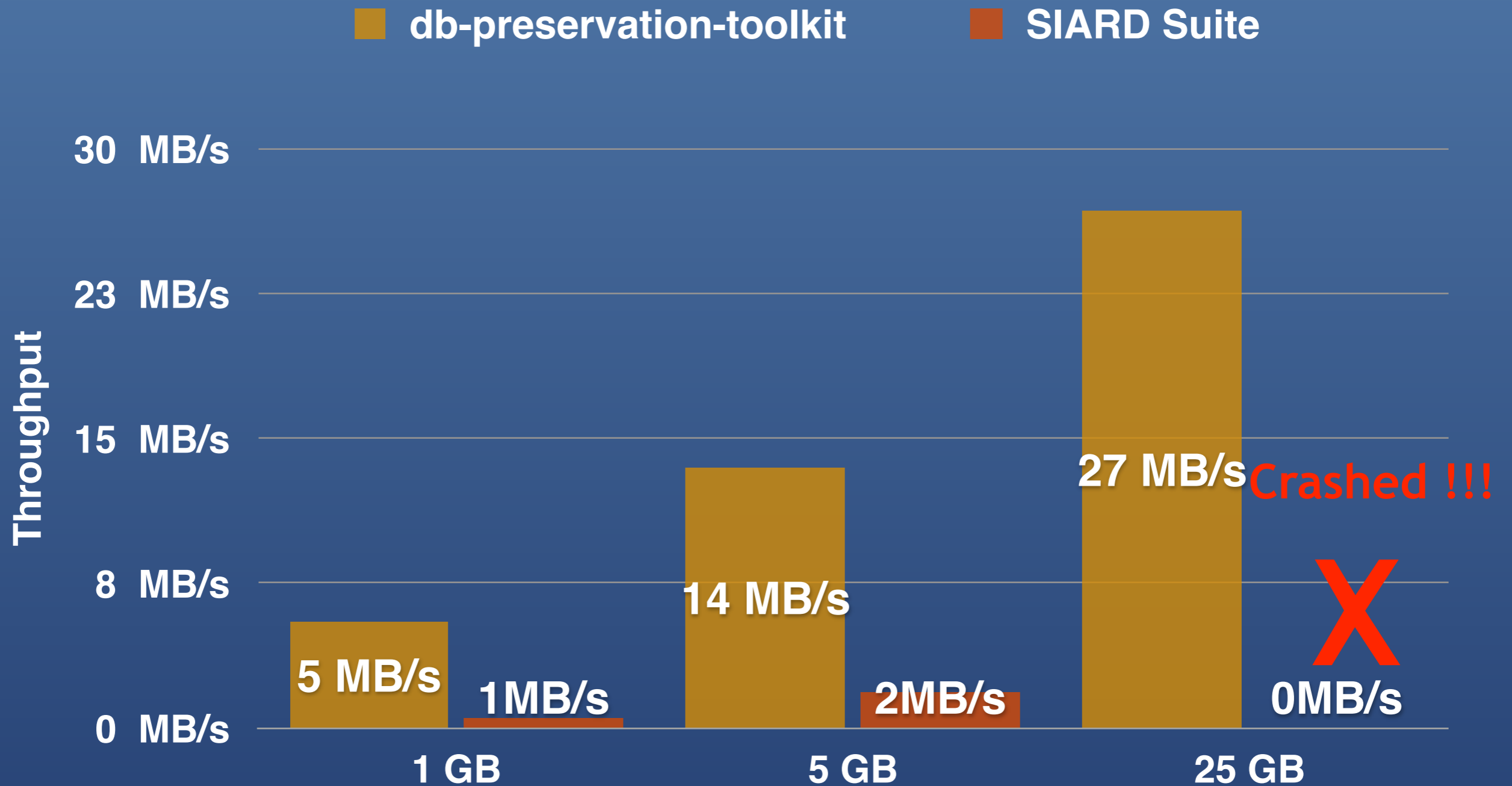
Textual dump + DBMS to SIARD conversion (white path)

Execute SIARD to DBMS conversion (red path) + Textual dump

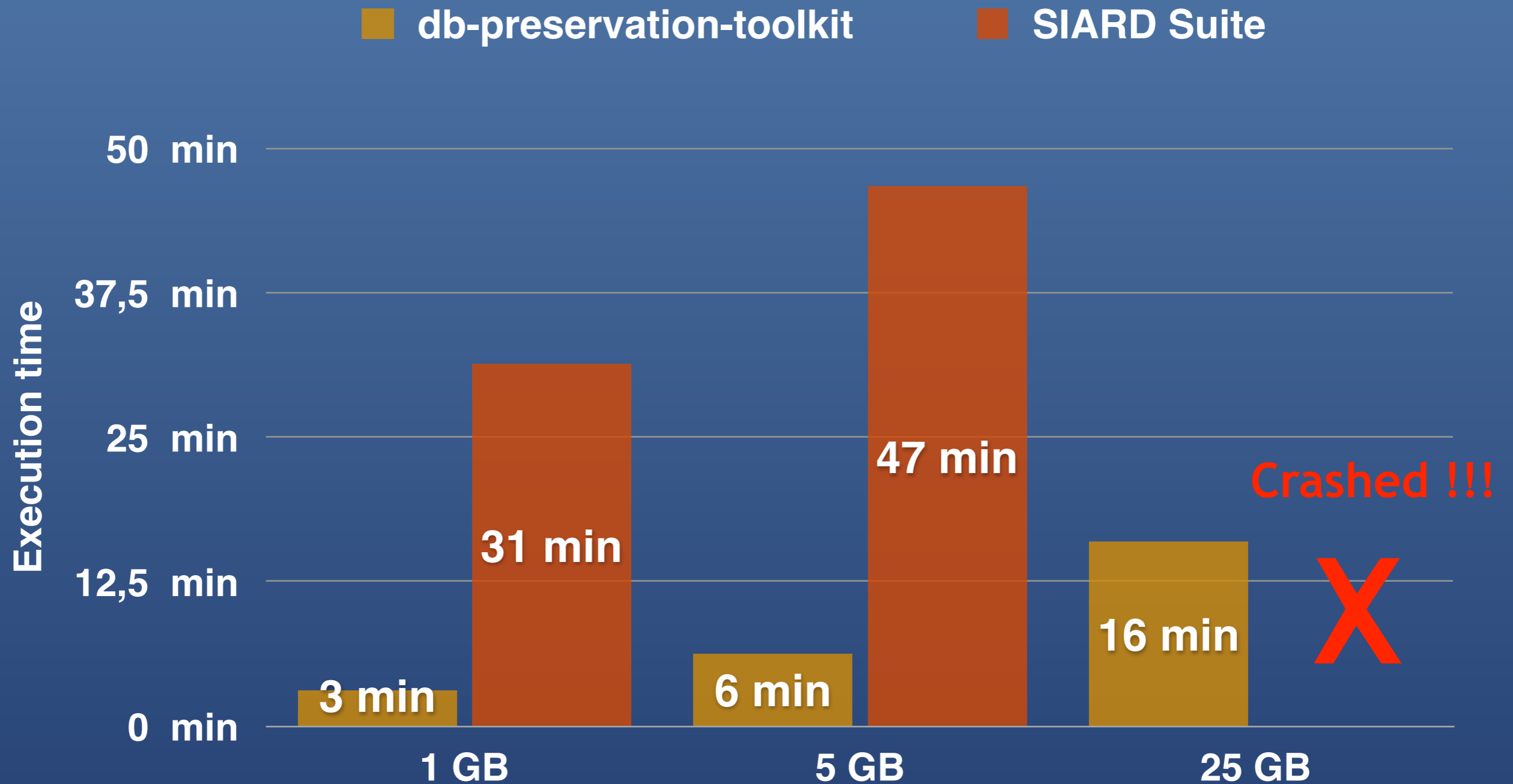
Do textual comparison of both dumps

Database cleanup (delete both databases created)

db-preservation-toolkit vs. SIARD Suite



db-preservation-toolkit vs. SIARD Suite



Future plans:
Fast database viewer

Preservation formats

DBML

SIARD

db-preservation-toolkit

Fast viewer resources

Lucene / Solr index

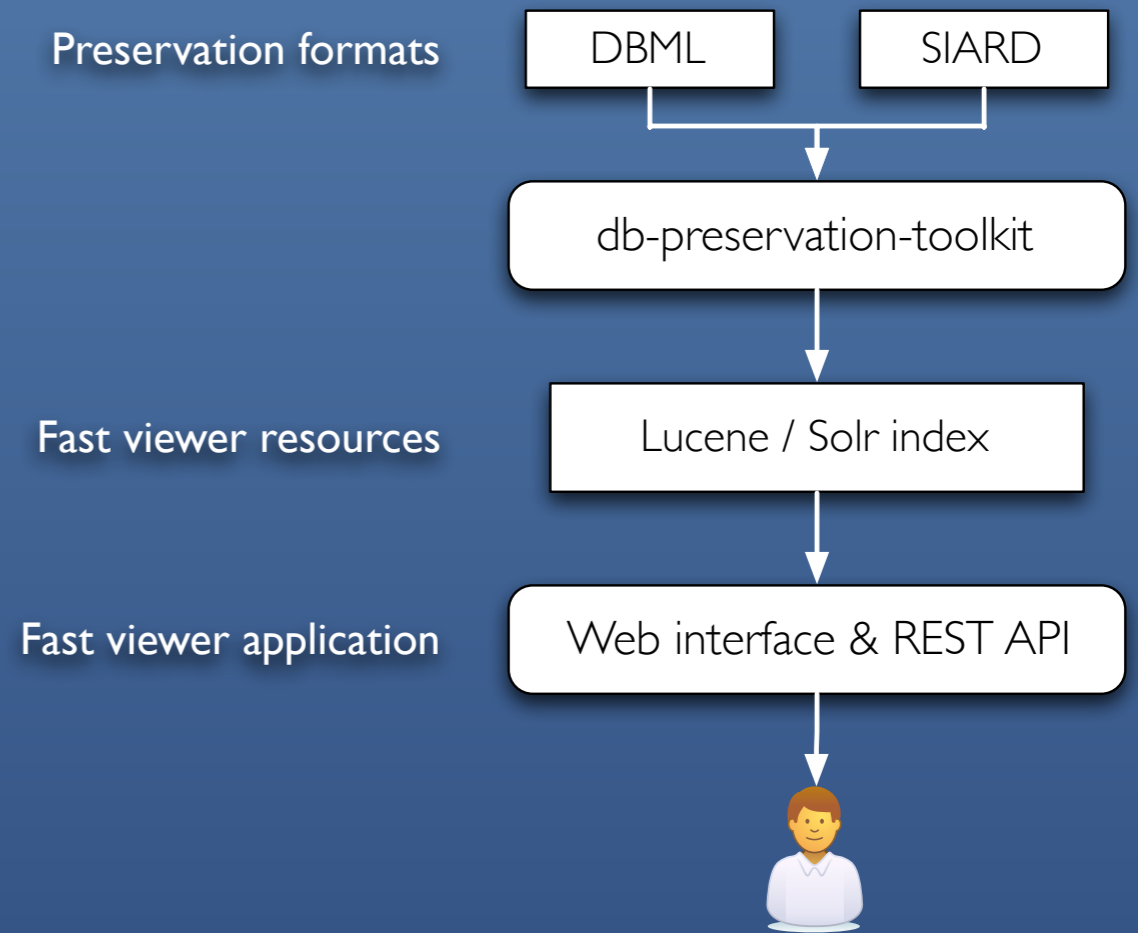
Fast viewer application

Web interface & REST API

Future plans:
Fast database viewer



Future plans: Fast database viewer



SIARD “flavours”

Danish (SIARD-DK)

E-ARK (SIARD-E/SIARD 2.0)

Database Preservation Toolkit

State-of-the-art and future plans
in the context of the E-ARK project

Is a Data Warehouse a Data Archive and Why Does It Matter?

Hélder Silva hsilva@keep.pt
KEEP SOLUTIONS

February 19-20, 2015