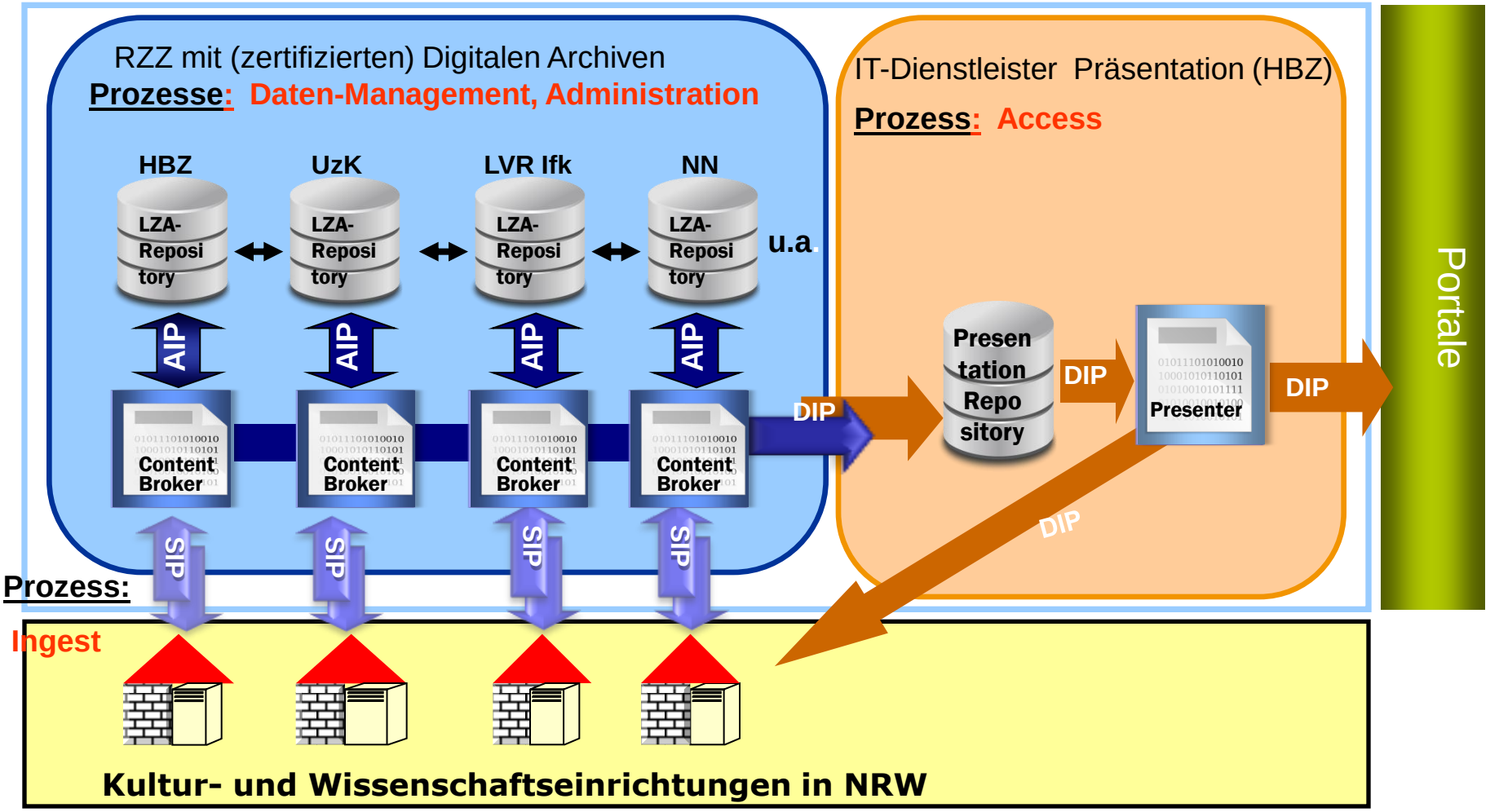
Historisch
Kulturwissenschaftliche
Informationsverarbeitung

DA-NRW: A distributed architecture for long-term preservation

Manfred Thaller, Sebastian Cuy, Jens Peters,
Daniel de Oliveira, Martin Fischer
Universität zu Köln

International Workshop on Semantic Digital Archives
Berlin, September 29th 2011

DA NRW: Systemarchitektur IT Verbund



SIP = Submission Information Packages
AIP = Archival Information Packages
DIP =Dissemination Information Packages

Basic (political) concepts 1 / 2

(1) Each cultural heritage institution in North Rhine Westphalia (NRW) can deposit digital material into the “Digital Archive” of the state.

(2) Such institutions are: Libraries, museums, archives, universities (publications as well as examination records), archaeological survey institutes, audiovisual archives …

(3) This includes a technology watch triggering migrations as necessary.

Basic (political) concepts 2 / 2

(4) The “Digital Archive” is a cluster of synchronized digital repositories at existing institutions. As a whole it implements OAIS.

(5) It also contains a “presentation component” which holds copies for unrestricted access via the large cultural heritage portals.

(6) Each depositor can decide, down to the item level, which policies are to be applied to the stored objects.

Technical purpose

(1) Prove of concept.

(2) Precise financial planning data for long term viability.

(3) For a restricted number of data formats and metadata standards.

(4) With a capacity of ca. 200 TB.

(5) No restrictions on scalability of number of formats and metadata standards.

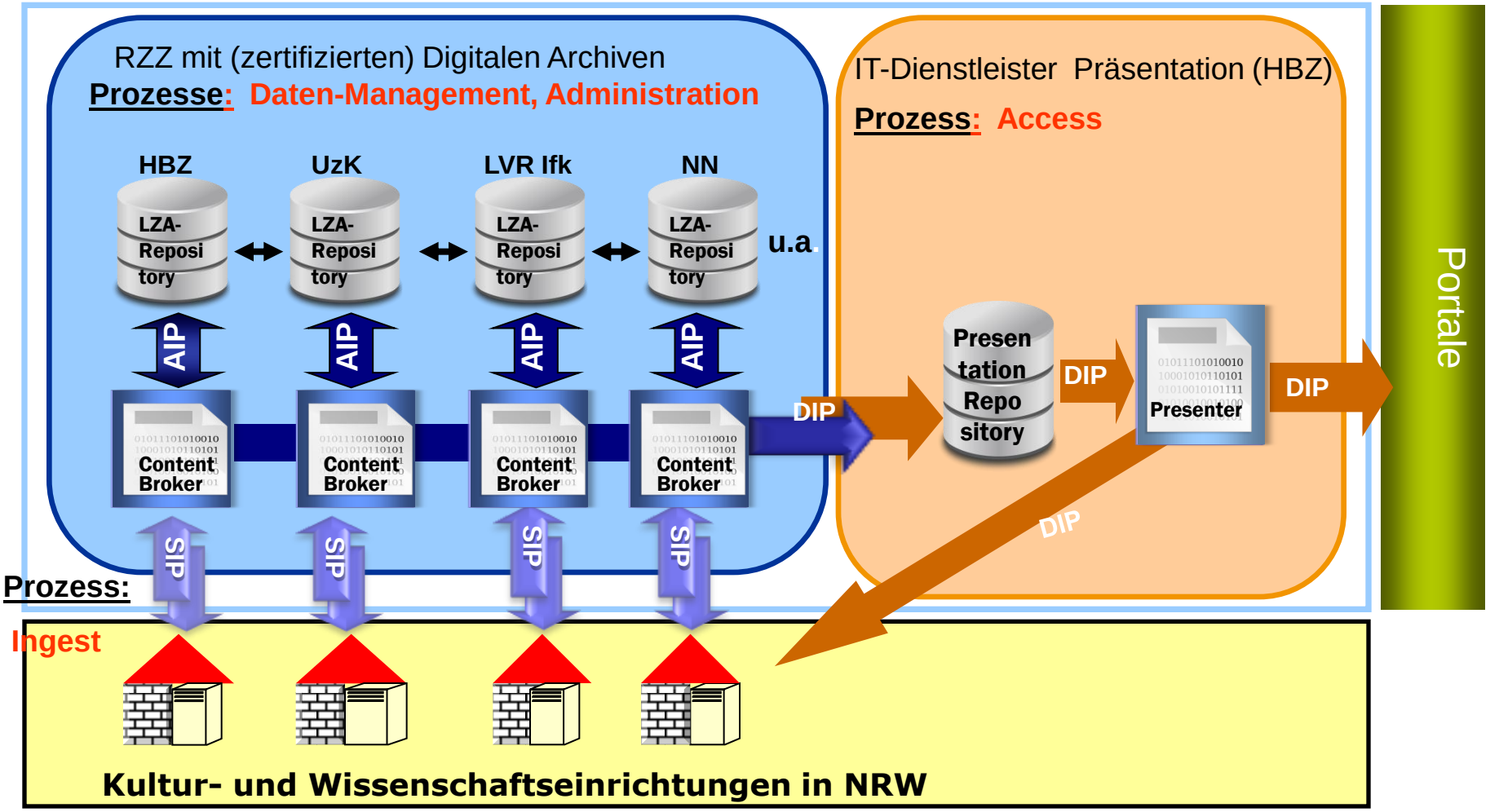
(6) Capacity scalable by at least one order of magnitude.

(7) Supposed to be operative in April 2012.

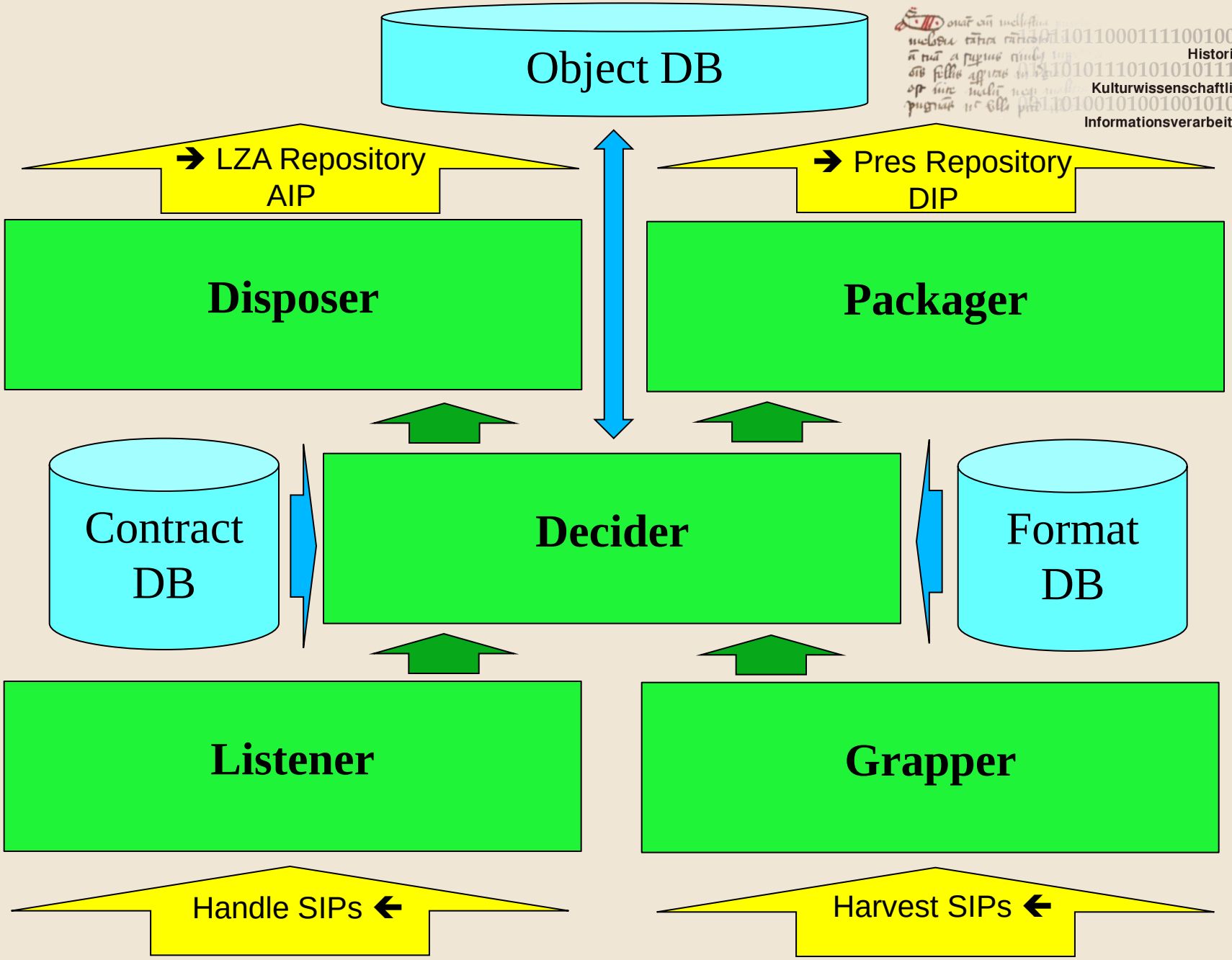
Small print

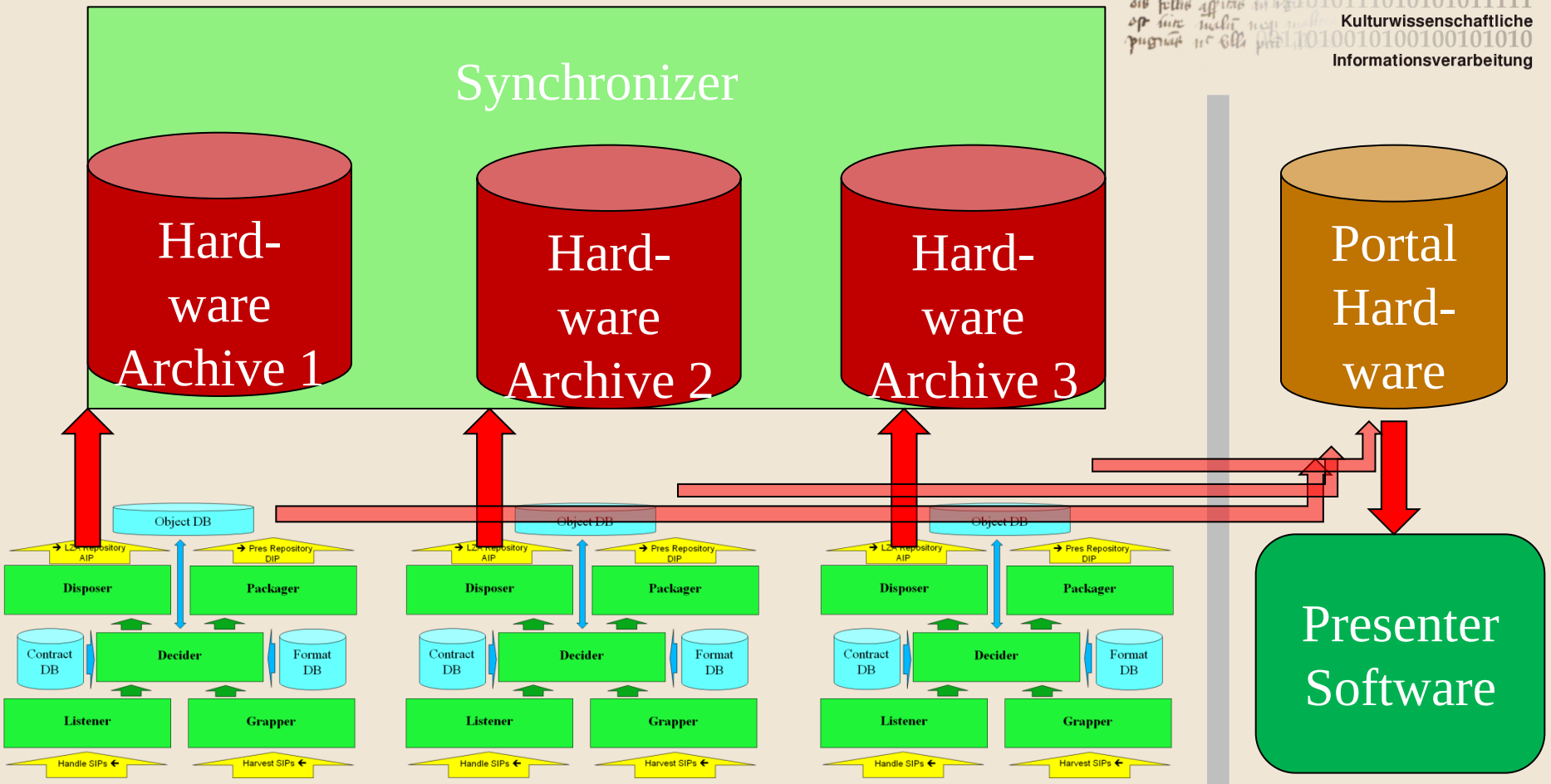
If the Eifel volcanoes explode, and only individual media survive, they should still be fully self descriptive.

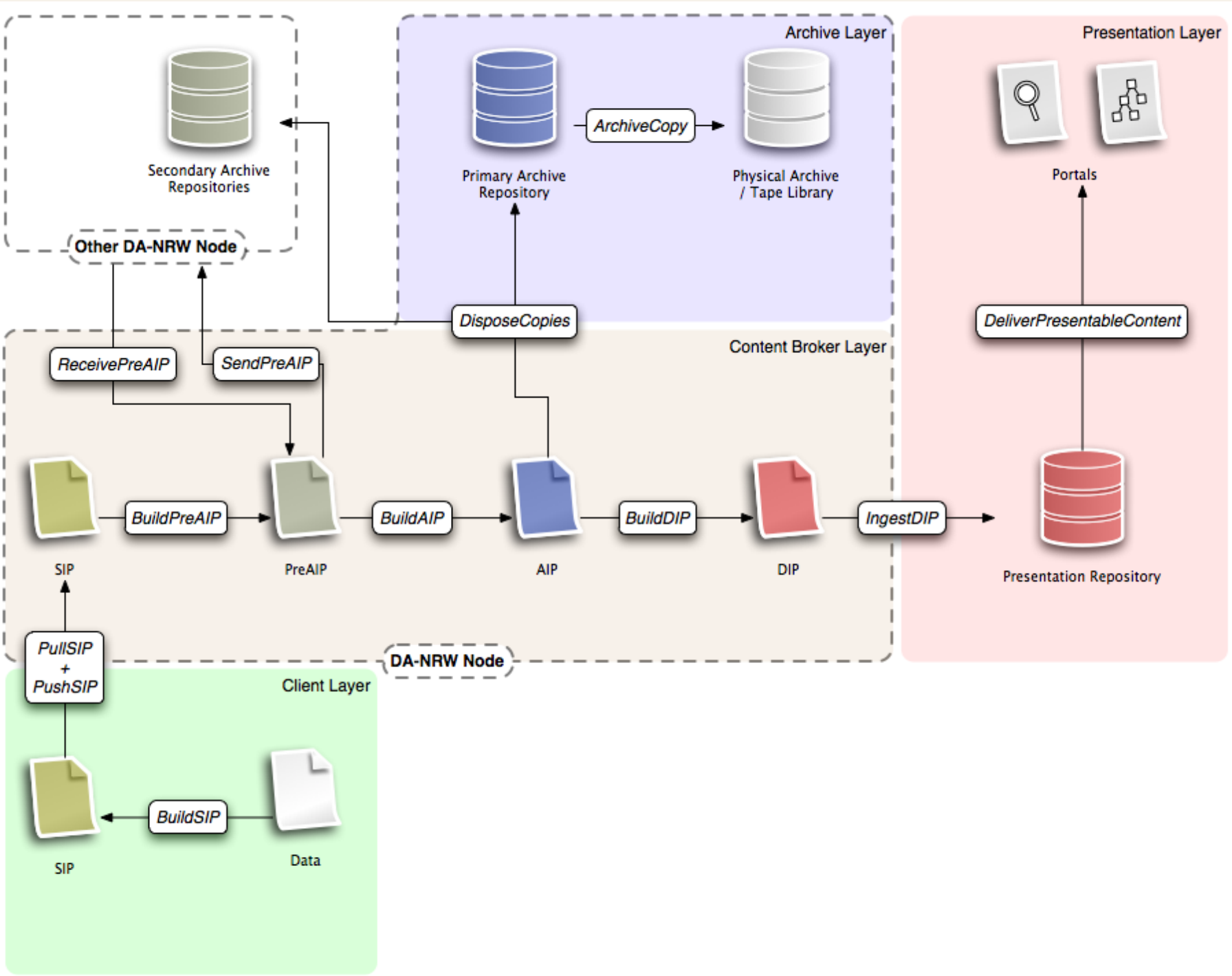
DA NRW: Systemarchitektur IT Verbund

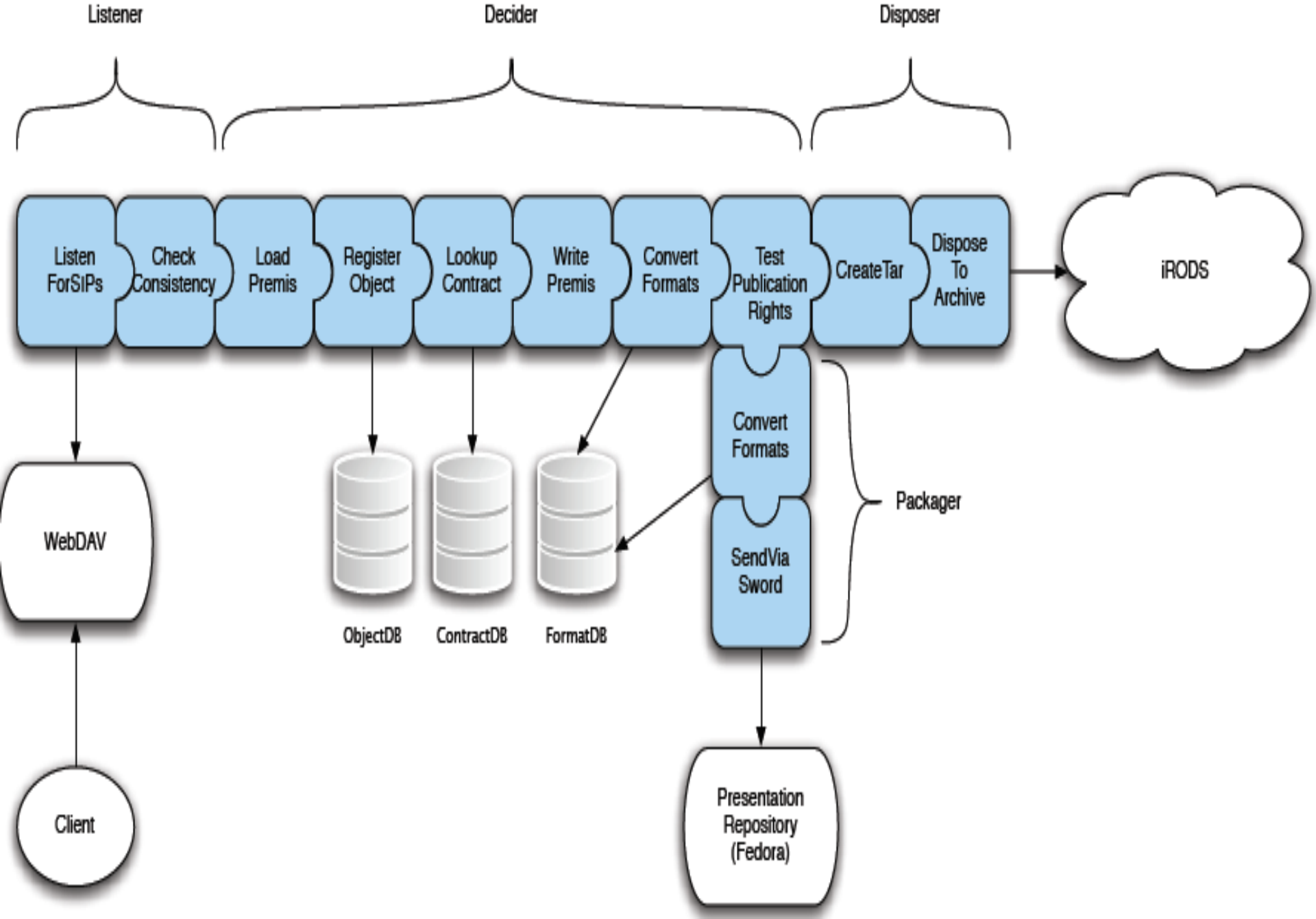


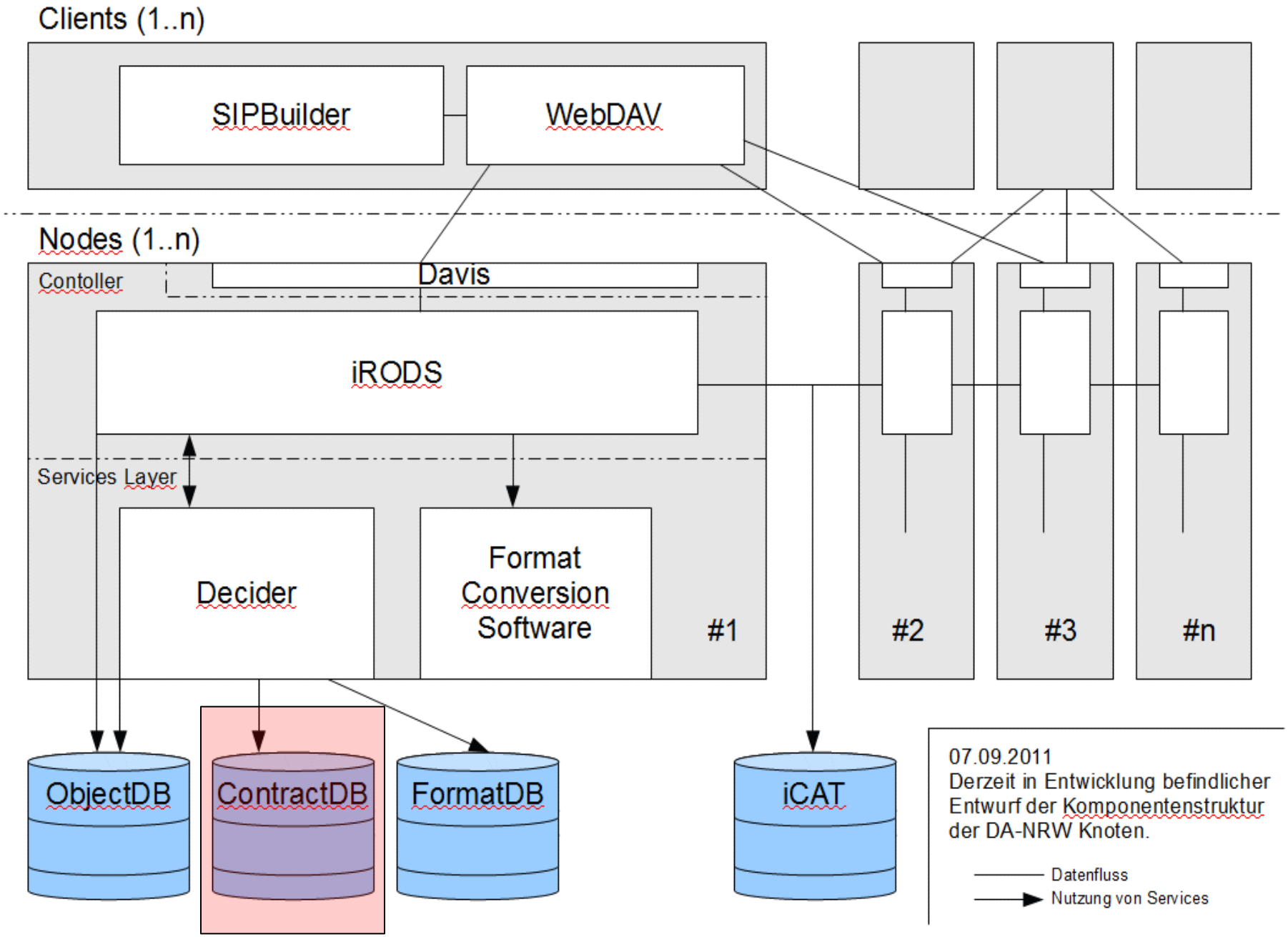
SIP = Submission Information Packages
AIP = Archival Information Packages
DIP =Dissemination Information Packages


Object DB
Historisch
Kulturwissenschaftliche
Informationsverarbeitung
➔ LZA Repository
AIP
➔ Pres Repository
DIP
Disposer
Packager
Contract
DB
Decider
Format
DB
Listener
Grapper
Handle SIPs ➔
Harvest SIPs ➔


Synchronizer
Hard-ware Archive 1
Hard-ware Archive 2
Hard-ware Archive 3
Portal Hard-ware
Object DB
Disposer
Packager
Contract DB
Decider
Format DB
Listener
Grapper
Handle SIPs
Harvest SIPs
Presenter Software


Archive Layer
Presentation Layer
Secondary Archive Repositories
Primary Archive Repository
ArchiveCopy
Physical Archive / Tape Library
Portals
Other DA-NRW Node
DisposeCopies
DeliverPresentableContent
Content Broker Layer
ReceivePreAIP
SendPreAIP
BuildPreAIP
BuildAIP
BuildDIP
IngestDIP
SIP
PreAIP
AIP
DIP
Presentation Repository
PullSIP + PushSIP
DA-NRW Node
Client Layer
BuildSIP
SIP
Data


Listener
Decider
Disposer
Listen ForSIPs
Check Consistency
Load Premis
Register Object
Lookup Contract
Write Premis
Convert Formats
Test Publication Rights
CreateTar
Dispose To Archive
iRODS
Convert Formats
SendVia Sword
Packager
WebDAV
ObjectDB
ContractDB
FormatDB
Client
Presentation Repository (Fedora)


Clients (1..n)
SIPBuilder
WebDAV
Nodes (1..n)
Contoller
Davis
iRODS
Services Layer
Decider
Format
Conversion
Software
#1
#2
#3
#n
ObjectDB
ContractDB
FormatDB
iCAT
07.09.2011
Derzeit in Entwicklung befindlicher
Entwurf der Komponentenstruktur
der DA-NRW Knoten.
Datenfluss
Nutzung von Services

Contracts 1 / 2

Generally speaking, an object will be kept on the repositories the system chooses and made available without restriction to the whole world.

A contract may:

- Restrict an object to specific repositories.
- Keep it out of the presentation system.
- Restrict the amount of metadata delivered to harvesters.
- Restrict the quality of data objects disclosed to harvester.
- …

Contracts 2 / 2

A contract may define these restrictions on:

- an institutional level.
- a media level.
- a object group level.
- the level of an individual object.

… out of which the concretely applicable contract for each archived object is computed.

Small print printed large

If the Eifel volcanoes explode, and only individual media survive, they should still be fully self descriptive.

Translated into:

- No separation of data and metadata, ever.
- Therefore use BagIt objects.

BagIt

Container format for AIPs:

- paket4223/
  - data/
    - bitstreams/
      - img2346.tif
      - img2347.tif
      - …
    - metadata/
      - dc.xml
      - premis.xml
      - mets.xml
      - …
  - manifest-md5.txt
  - bagit.txt

```
<rightsExtension>
  <ar xmlns="http://danrw.de/access" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" schemaLocation="http://danrw.de/access file:/Da-NRW/RechtemanagementAKI/PremisAbbildung/ar.xsd">
    <conditions>
      <PUBLICATION >
        <ip_range required="false"> <val>44.56.81.211</val> <val>64.56.81.211</val> </ip>
        <displayResolutionImage> <width_px>200</width_px> </displayResolutionImage>
        <streamingQuality> <kbps>64</kbps> </streamingQuality>
        <pageLimit> <pages>2</pages> </pageLimit>
      </PUBLICATION>
    </conditions>
  </ar>
</rightsExtension>
```

BagIt / Premis extended

Premis within BagIt extended by contractual information stored within objects in OWL … and OWL representation of contracts handled as contract DB within a triplestore.

Probably.

Technology decisions

(1) Own uppermost orchestration layer …

(2) … supported by iRods micro service layer

https://www.irods.org

(3) No Fedora, when performance counts. Use inpresentation level, however.

http://fedora-commons.org

Standards to be used

(1) LTP objects: BagIT http://tools.ietf.org/html/draft-kunze-bagit-06

(2) object identifiers: URN

(3) Structural metadata: METS / MODS, as far as applicable; otherwise domain specific equivalent

(4) [format identifiers: PRONOM / UDFR version, RDF based]

(5) PREMIS: PREMIS. Currently XML binding, *probably* RDF or UML binding

Why probably?

If the Eifel volcanoes explode, and only individual media survive, they should still be fully self descriptive.

Do Semantic Technologies make long term preservation easier or harder?

Gladney: *Digital Archiving* v. *Long Term Preservation*

Historisch
Kulturwissenschaftliche
Informationsverarbeitung

Thank you for listening!

sebastian.cuy@uni-koeln.de
d.de-oliveira@uni-koeln.de
martin.fischer@uni-koeln.de
jens.peters@uni-koeln.de
manfred.thaller@uni-koeln.de