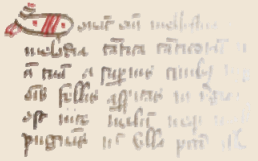
Digitales Archiv NRW

Manfred Thaller
Universität zu Köln

Hamburg, DINI: Elektronische Langzeitarchivierung in Hochschulen,

29. Februar 2012



I. Vorgaben

Langfristiges Ziel / Inhalt 1

(1) Jede Kulturerbeeinrichtung in NRW soll die Möglichkeit haben beliebigen digitalen Inhalt an ein DANRW digital (über das Netz) *abzuliefern*.

(2) Dies führt dazu, dass dieser Inhalt in einem OAIS kompatiblen landesweiten Langzeitarchiv abgelegt wird.

(3) Dies schließt eine Technology Watch und Migrationen bei Bedarf ein.

Langfristiges Ziel / Inhalt 2

(4) Jede Kulturerbeeinrichtung in NRW kann bestimmen, in welchem Umfang der archivierte Inhalt an Portale weitergegeben werden darf.

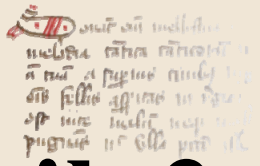
(5) Nach Maßgabe dieser Regelungen werden aus den abgelieferten Daten und Metadaten geeignete Darstellungen für DDB und Europeana abgeleitet, die via OAI PMH und anderer Protokolle auch allgemein bereitstehen.

Langfristiges Ziel / Technik 1

(1) *Existierende* IT Infrastruktureinrichtungen im Lande NRW stellen drei oder mehr „Knoten“ aus Rechner und Speicherkapazität bereit.

(2) Diese erscheinen den Nutzern als ein einheitliches Langzeitarchiv für das Land.

(3) Die drei+ Knoten synchronisieren sich selbständig und gleichen die Daten regelmässig gegeneinander ab.



Langfristiges Ziel / Technik 2

(4) „Eifelvulkanprinzip“: Wird nach der Eruption der Eifelvulkane ein Datenträger gefunden, enthält er vollständige, verarbeitbare digitale Objekte.

Kurzfristiges Ziel = Vorprojekt

(1) Nachweis der technischen Möglichkeit.

(2) Exakte Kostenschätzung für den Dauerbetrieb.

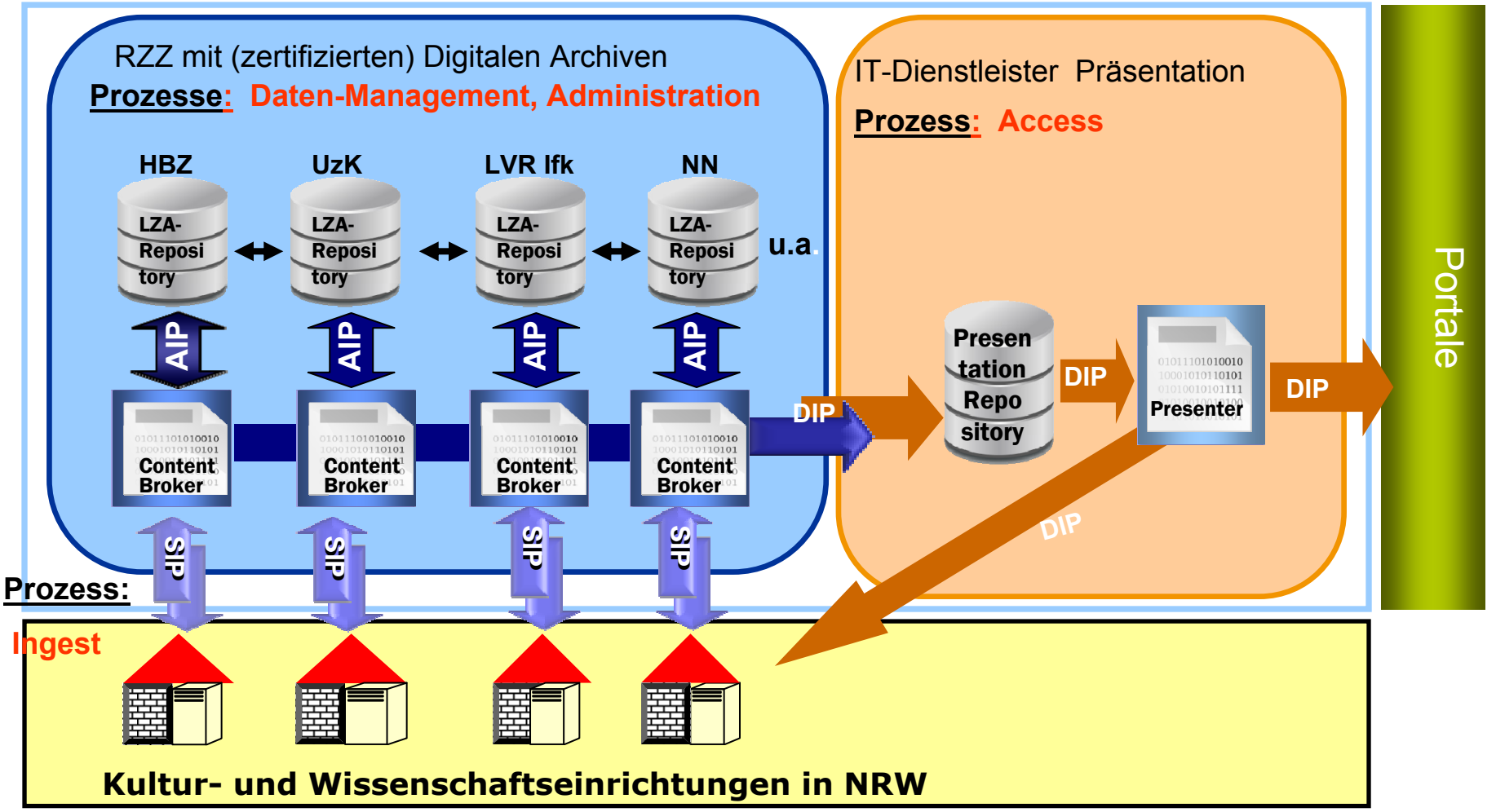
(3) Für eine beschränkte Zahl von Datenformaten und Metadatenstandards.

(4) Bei einer Kapazität von etwa 200 TB.

(5) Bei Formaten und Metadatenstandards unbeschränkt skalierbar.

(6) Kapazität skalierbar um mindestens eine Größenordnung.

DA NRW: Systemarchitektur IT Verbund



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

“Konsortium” 1 / 2

Kulturinstitutionen:

Theater der Klänge

Kunst- und Museumsbibliothek Köln

Landesarchiv NRW

LVR Archivberatungs- und Fortbildungszentrum

LVR Industriemuseum Oberhausen

LVR Zentrum für Medien und Bildung

LWL Medienzentrum

Universitäts- und Landesbibliothek Bonn,

Universitäts- und Landesbibliothek Düsseldorf

Universitäts- und Landesbibliothek Münster

“Konsortium” 2 / 2

Koordination:

Ministerium für Familie, Kinder, Jugend, Kultur, Sport des Landes Nordrhein-Westfalen

Technik:

Universität Köln, Hist.-kulturwissenschaftliche Informationsverarbeitung (HKI)

Universität Köln, Regionales Rechenzentrum (RRZK)

Hochschulbibliothekszentrum NRW (hbz)

Landschaftsverband Rheinland, LVR-Infokom

HKI

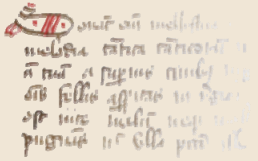
„Fachinformatik der Geisteswissenschaften“

Beschäftigung mit Langzeitarchivierung:

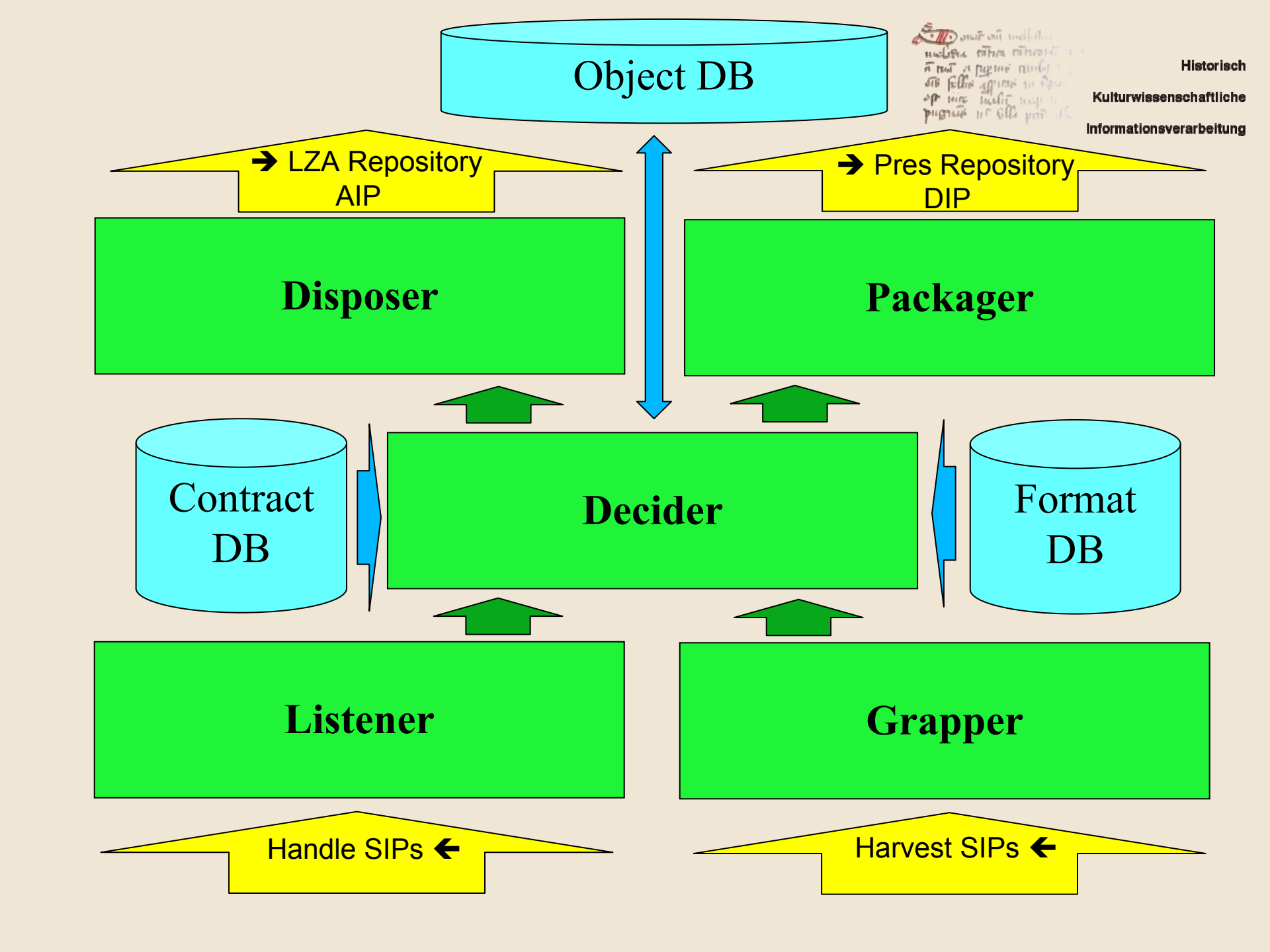
- EU Projekte, u.a. Delos (http://www.delos.info/) und Planets (http://www.planets-project.eu/; dazu auch http://planetarium.hki.uni-koeln.de/planets_cms/index.php).
- Lehre auf ca. 12 LZA Summerschools seit 2004.
- Kooperation mit http://www.openplanetsfoundation.org/

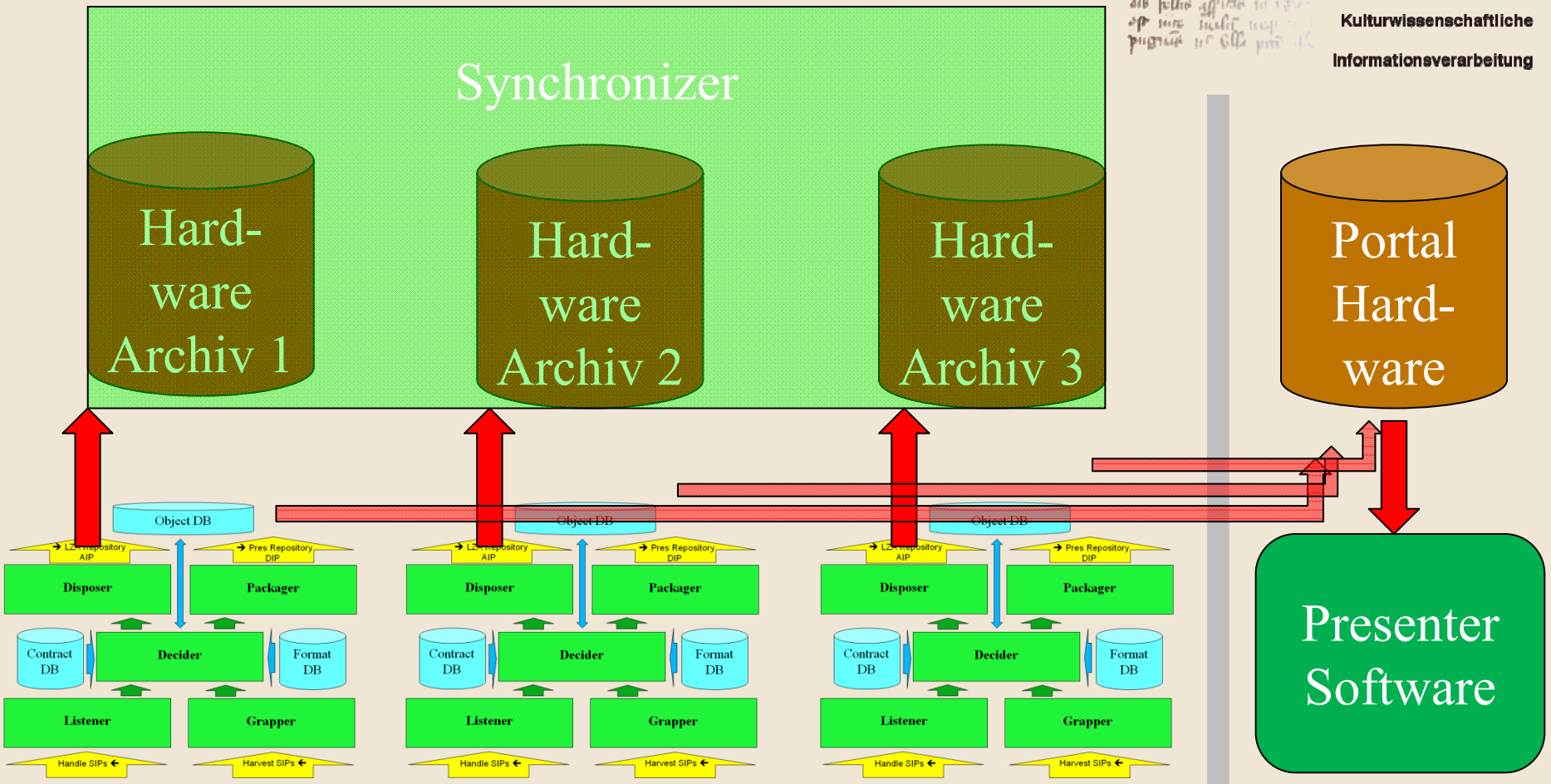
Zum DANRW:

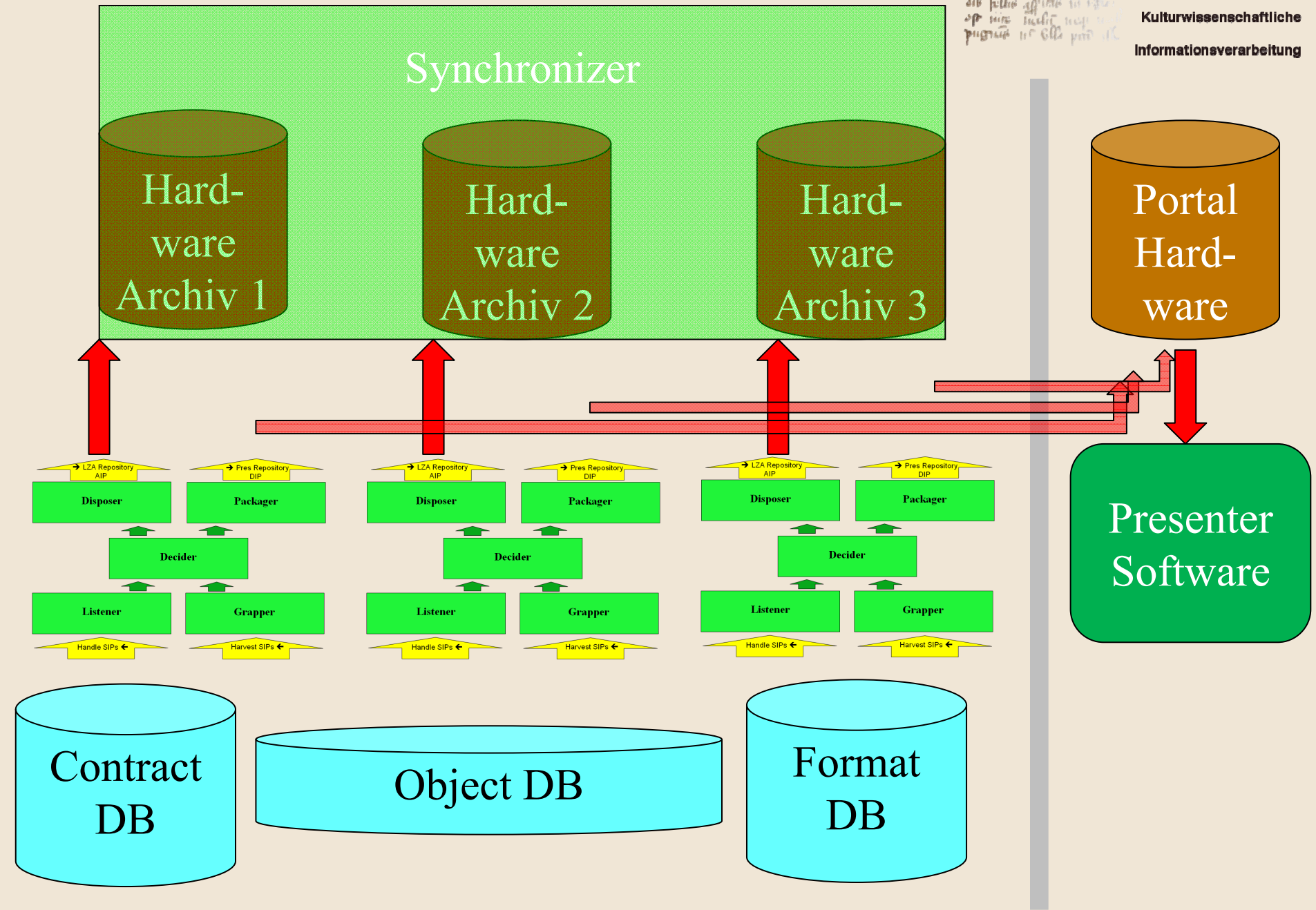
M.Thaller et al: „DA-NRW: a distributed architecture for long-term preservation”, in: L. Predoiu et al (Hg.): *Semantic Digital Archives*, Berlin, 2011, (http://ceur-ws.org/Vol-801/paper13.pdf)

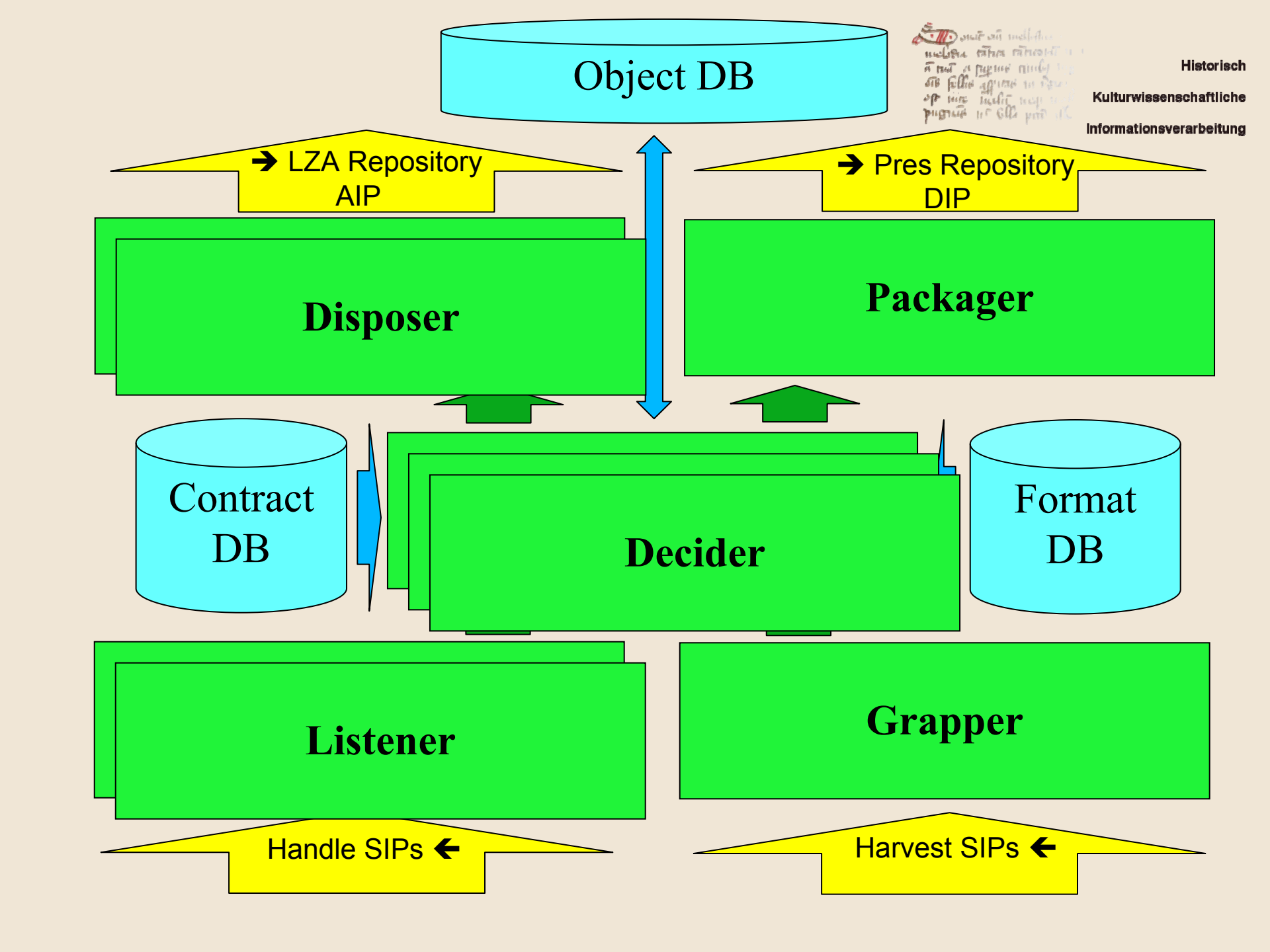


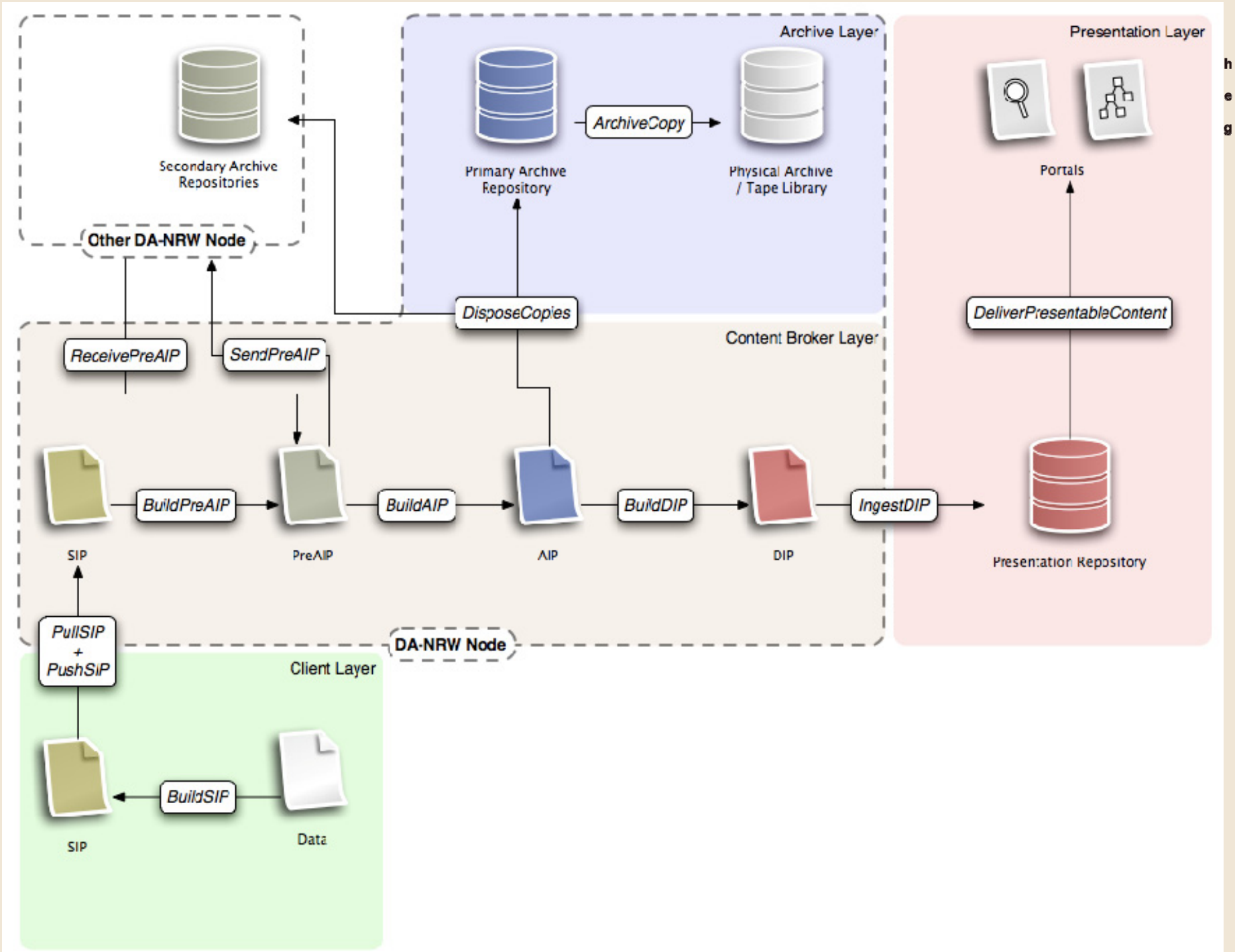
II. Basisarchitektur

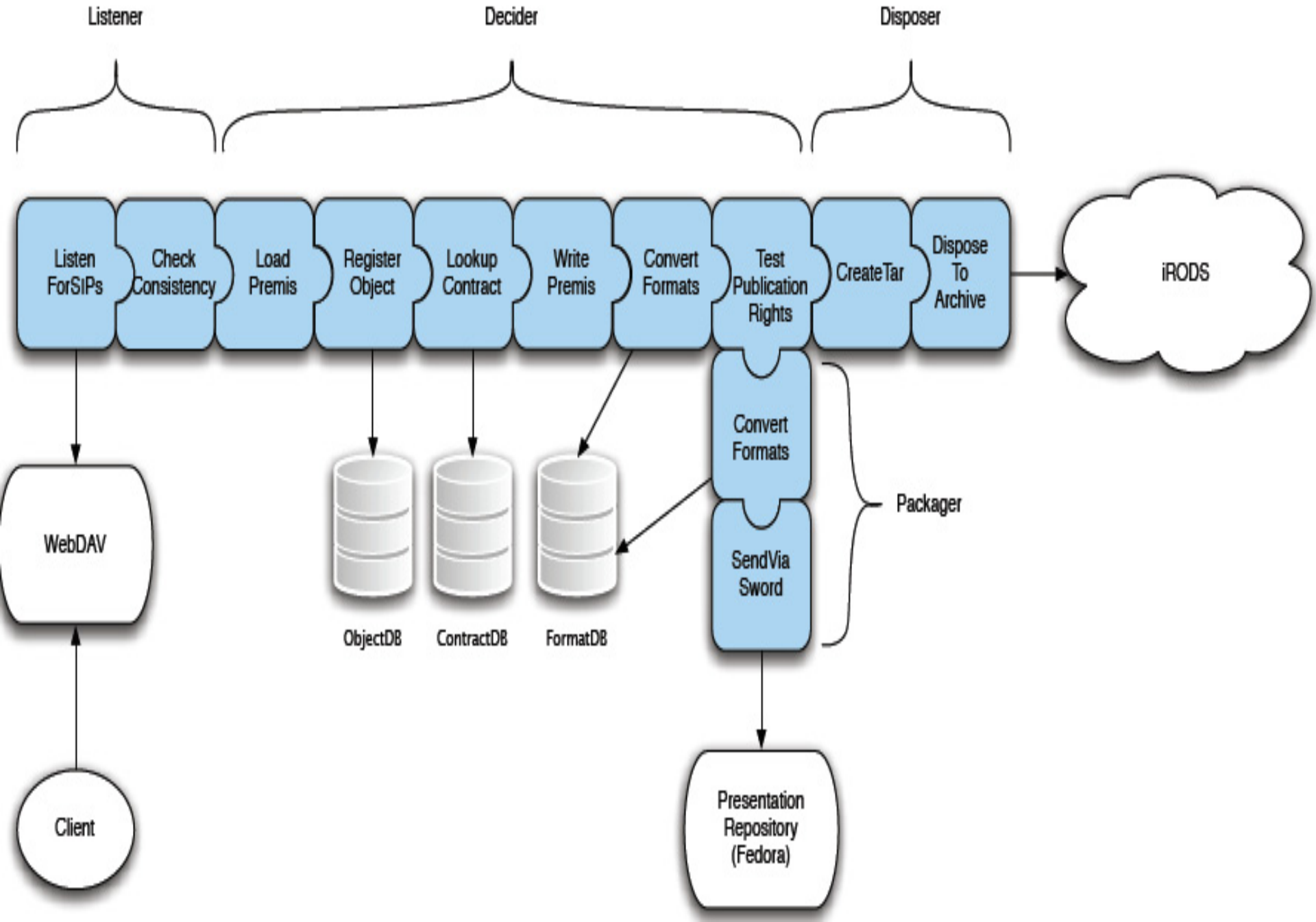

Historisch
Kulturwissenschaftliche
Informationsverarbeitung
Object DB
➔ LZA Repository
AIP
➔ Pres Repository
DIP
Disposer
Packager
Contract
DB
Decider
Format
DB
Listener
Grapper
Handle SIPs ➔
Harvest SIPs ➔


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


Synchronizer
Hard-ware Archiv 1
Hard-ware Archiv 2
Hard-ware Archiv 3
Portal Hard-ware
Presenter Software
LZA Repository AIP
Pres Repository DIP
Disposer
Packager
Decider
Listener
Grapper
Handle SIPs
Harvest SIPs
Contract DB
Object DB
Format DB


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


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)

III. Umsetzung eines inhaltlichen Konzepts in einen Use Case

Beispiel: *Contracts*

Konzept des “Contracts”

Eine Beschreibung der Bedingungen, unter denen eine Einrichtung dem DA NRW digitale Inhalte zur LZA übergibt.

(1) Gespeichert als XML Dokument.

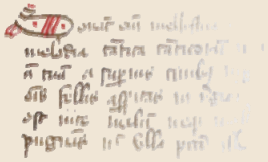
(2) Verfügbar als gedrucktes und ggf. gezeichnetes Dokument.

“Bedingungen”

(1) Das Land NRW stellt eine Infrastruktur bereit, die sicherstellt, dass das digitale Kulturerbe des Landes auf Dauer sicher ist.

(2) Das Land NRW stellt eine Infrastruktur bereit, die sicherstellt, dass das digitale Kulturerbe des Landes international die größtmögliche Sichtbarkeit hat.

(3) Ausnahmen sind möglich.

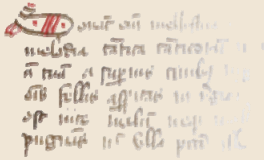


Inhalt des “Contracts”

Einschränkungen bei der Benutzung von Ressourcen innerhalb des DA NRW.

Einschränkungen der an den Presenter weiterreichbaren Metadaten.

Maximale Qualität der an den Presenter weiterreichbaren Metadaten.



Ebenen des “Contracts”

Ablieferungen einer Institution.

Ablieferungen eines Datentyps.

Ablieferungen einer Objektgruppe.

Ablieferung eines einzelnen Objekts.

➔ Resultierender „Object contract“ wird Teil der LZA Metadaten.

IV. Umsetzung einer organisatorischen Annahme in konkrete Komponenten

Beispiel: *Pre-Ingest*

Beispiel Ingest: Prämissen 1 / 2

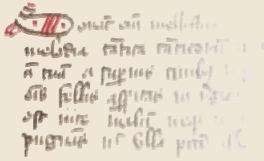
Grundsatz DA NRW: *Übernahme* von SIPs nach DIN 31645

Beispiel: (Standardentwurf 5.2.4 – “Zeitliche Abnahme der Validierungsprozesse”)

In einer ersten Phase werden Kriterien abgefragt, die vollständig erfüllt sein müssen:

a. Enthält die Lieferung alle vereinbarten digitalen Objekte?
b. Sind die digitalen Objekte integer?

Beispiel Ingest: Prämissen 2 / 2



Grundsatz DA NRW: *Übernahme* von SIPs nach DIN 31645

Beispiel: Vorgesehene Umsetzung DA-NRW seitig

a. Sind alle in den SIP Metadaten erwähnten physikalischen Dateneinheiten vorhanden?
b. Werden alle vorhandenen Dateneinheiten in den Metadaten erwähnt?
c. Sind die Checksummen aller Dateneinheiten verifizierbar?
d. Sind alle *supported* und *understood* formats „validierbar“?

Beispiel Ingest: Konsequenz

(1) Es wird erwartet, dass die abliefernden Einrichtungen in der Lage sind, korrekte SIPs abzuliefern.

(2) Deshalb wurde eine Vereinbarung mit einem Service Provider getroffen, wonach dessen Ablieferungsformat akzeptiert wird.

(3) Es wird aber auch ein *SIP Builder* angeboten, der Daten und Metadaten zusammenfasst, mit Checksummen versieht und für den Transfer vorbereitet.

V. Technologieentscheidungen / gewählte Standards

Technologieentscheidungen

(1) Orchestrierung durch eigene Module oder als Teil …

(2) hardwarenaher Komponenten auf der Basis von iRods. (iRods / University of North Carolina at Chapel Hill: Basis für NARA Transcontinental Persistent Archive Prototype.)

https://www.irods.org

(3) Presentation Repository auf der Basis von Fedora.

http://fedora-commons.org

Zentrale Standards

(1) OAIS konform … als Kern

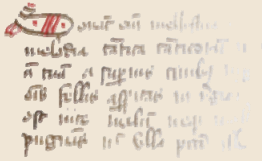
(2) LZA Objekte: BagIT http://tools.ietf.org/html/draft-kunze-bagit-06

(2) Objektidentifikatoren: URN

(3) Strukturdaten: METS / MODS, so anwendbar

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

(5) PREMIS: PREMIS. Derzeit XML Bindung, kann noch durch RDF oder UML Bindung ersetzt werden



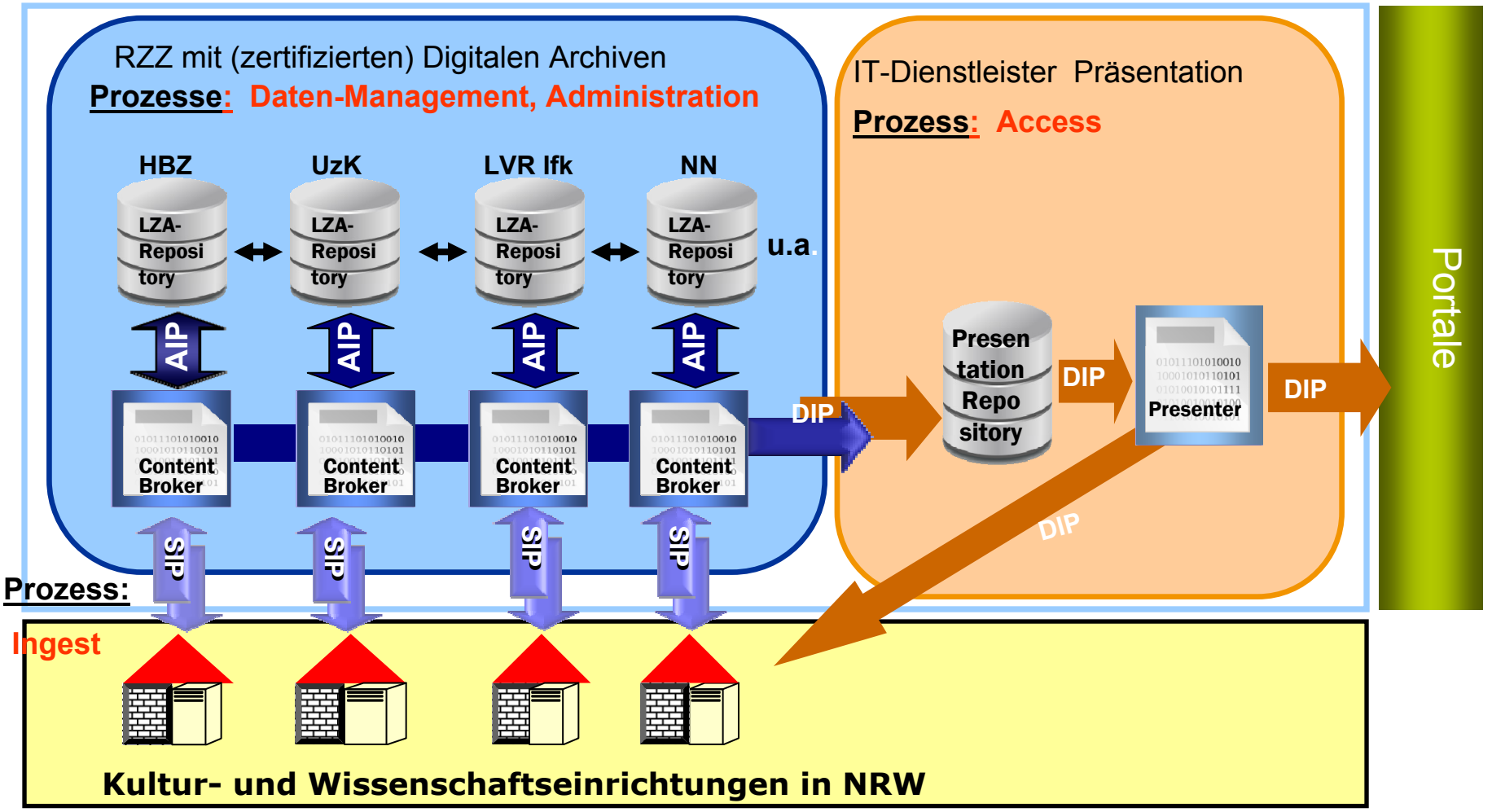
VI. Status

Projekt DANRW / Vorprojekt - Stand

Ziel: „Erweiterte Machbarkeitstudie“

Zeitplan:
(1)Organisatorische Vorbereitungen = Nov. / Dez. 2010
(2)Vorphase: Design = Jan. / Feb. 2011
(3)Iteration I: Pre-Prototyp = März / Juni 2011
(4)Iteration II: Stabiler Prototyp = Juli / Dez. 2011
(5)Iteration III: Skalierbares System = Jan. / April 2012
(6)[Verstetigung = Mai / Dezember 2012]

DA NRW: Systemarchitektur IT Verbund



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

Projekt DANRW / Vorprojekt – Stand 1

Ziel: „Erweiterte Machbarkeitstudie“

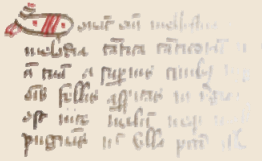
Status 29. Februar 2012:
„Kontrollierter Produktionsbetrieb“ auf der Basis der ULB Bestände
Tests auf Basis der Bestände anderer Partner

Projekt DANRW / Vorprojekt – Stand 2

Ziel: „Erweiterte Machbarkeitstudie“

Mutmaßlicher Status 30. April 2012:
LZA – „Beschränkter Produktionsbetrieb“ 30 – 40 TB
Portal – DDB ja; sonst: Wiedervorlage

- Technologywatch: Schnittstellen.
- „Eifelvulkanprinzip“ noch nicht umgesetzt.



VII. Lehren

Meinungen & Erfahrungen

Claire Bretecher:
http://www.clairebretecher.com/NET/classic/casanova.htm

Meinungen & Erfahrungen

Langzeitarchivierung jetzt!

Meinungen & Erfahrungen

Bitstream preservation alleine reicht nicht; ohne Bitstreams sind die Metadaten aber nutzlos.

Meinungen & Erfahrungen

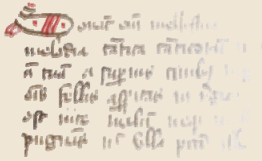
Ein Preservation Strategie sollte mit dem Allgemeinen beginnen, die Ausnahmefälle sollten DANACH diskutiert werden.

Meinungen & Erfahrungen

Agile Methodologies sind praktisch sehr relevant. Der "Wasserfall" entspricht NICHT mehr dem State of the Art.

Meinungen & Erfahrungen

Digitale LZA kann organisch in bestehende Infrastrukturen integriert werden und erfordert keine grundlegenden Neukonstruktionen. Nachdem sie ein erhebliches *operatives* Knowhow erfordert, hat dies auch generell positive Aspekte.



Herzlichen Dank!

manfred.thaller@uni-koeln.de