

DIE ÖSTERREICHISCHE BIBLIOTHEKENVERBUND
UND SERVICE GMBH



obv sg

Inkrementelle Datenexporte aus dem Österreichischen Bibliothekenverbund

Probleme und Lösungen

V. Babitchev – J. Brandauer

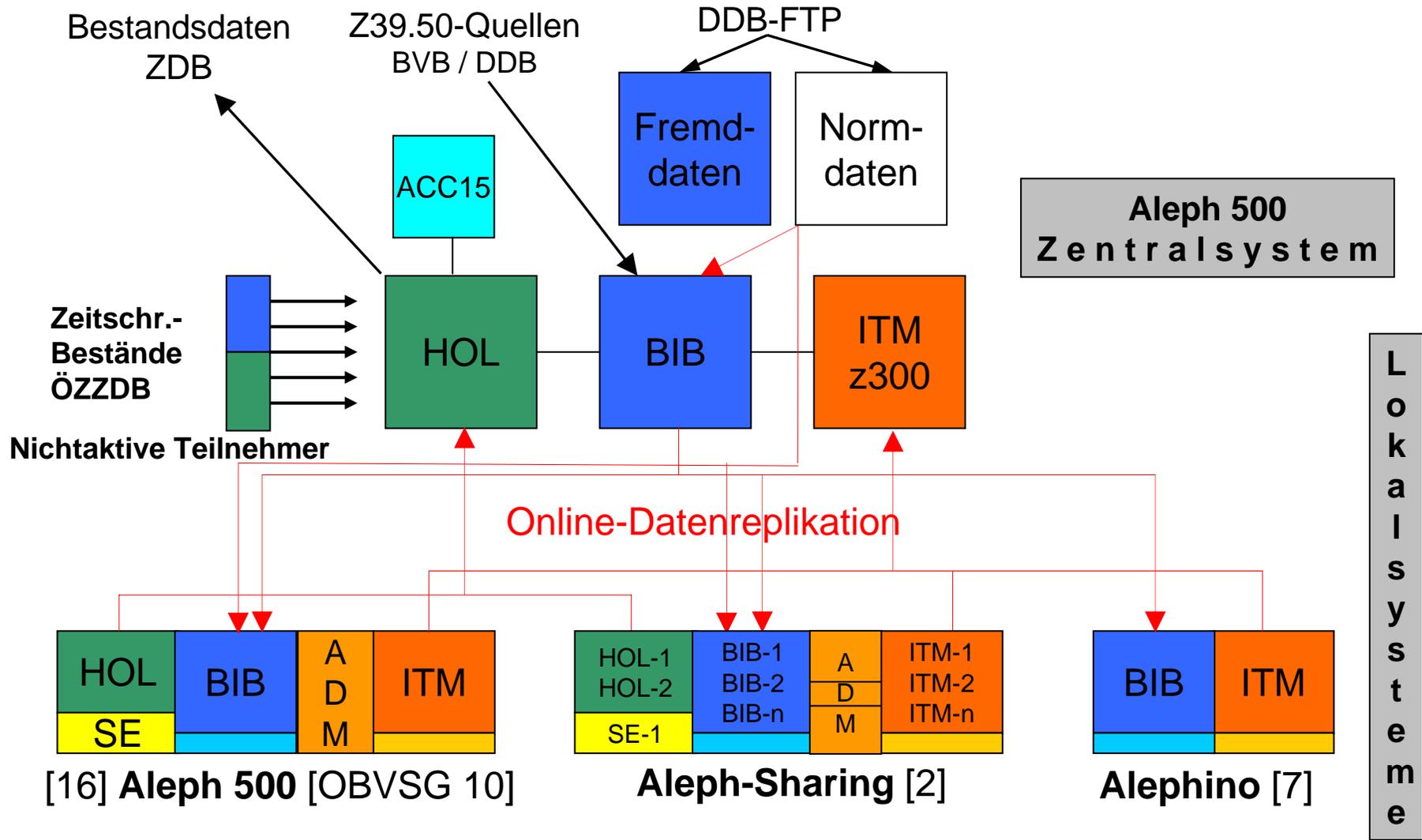
Aleph-DACH-Tagung 2006 – ETH Zürich – 21.-22.August 2006

Übersicht

Übersicht

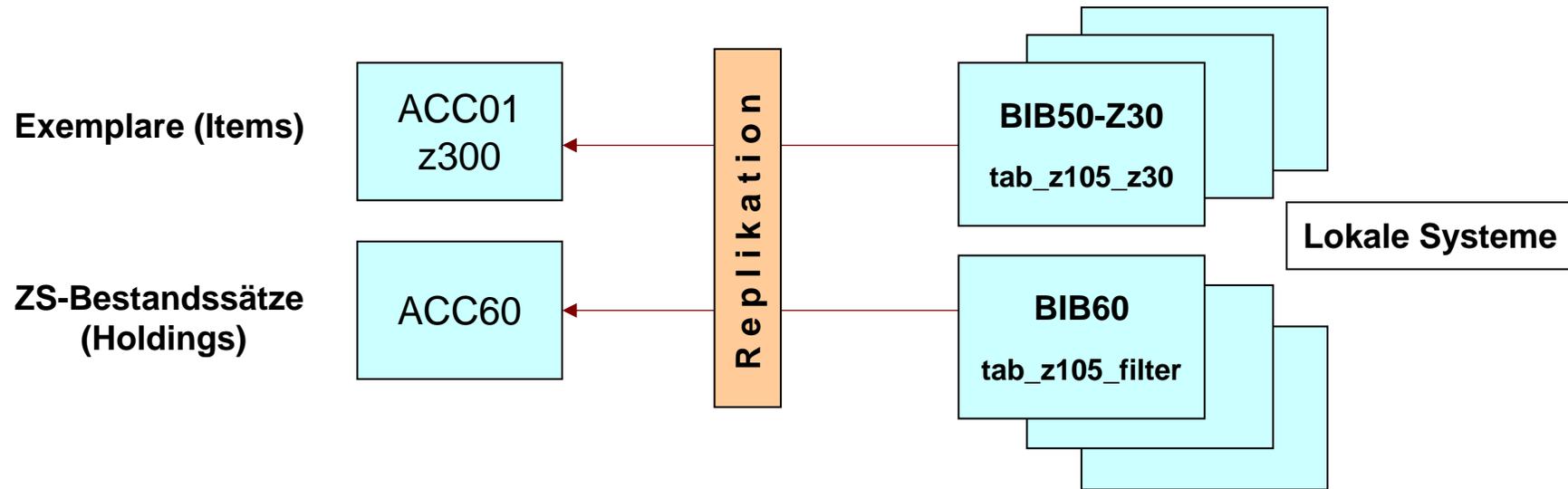
- Verbundstruktur des OBV
- Speicherung der Lokaldaten im Verbundsystem
- Projekte mit Datenexporten
- Anforderungen an Projektteilnehmer
- Registrierung von BIB- und HOL-Änderungen
- Registrierung von BIB- und HOL-Löschungen
- Registrierung von z300 Änderungen / Löschungen
- Export der Lokaldaten
- Workflow Export Suchmaschinen
- Zusammenfassung

Verbundstruktur des OBV



Speicherung der Lokaldaten im Verbundsystem

Speicherung der Lokaldaten im Verbundsystem



- Jeder Verbundteilnehmer verfügt über (teilt mit anderen) ein lokales System
- keine Lokalsysteme im Verbund!
- **Exemplare – z300:** Keine Replikation vom Typ ISSBD und ISSUE
- Z300 entspricht komprimierter Z30 (alle Informationen vorhanden)
- **ZS-Bestandssätze - HOL:** Keine Replikation lokaler Sacherschliessungs-Sätze

Projekte mit Datenexporten

Versorgung von Suchmaschinen

Ziel:

Erprobung von Suchmaschinen als alternative zum bestehenden Web-Opac sowie Bereitstellung zusätzlicher Recherche-Zugänge

- Dreiländerkatalog – hbz [FAST]
- eDOC – OBSG [Swish-e]
- Google Scholar

Versorgung von externen Katalogen

Ziel:

Datenpräsenz in überregionalen Katalogen sowie Teilnahme an Kooperationen

- Österreichische Zeitschriftenbestände in ZDB
- Projekt kooperative Neukatalogisierung

Anforderungen an Projektteilnehmer

Anforderungen an die Projektteilnehmer

	Daten			Export-Typ		Format	Anmerkung zum Format
	BIB	HOL	ITM	VOLL	INKR		
Dreiländerkatalog	X	X	X	X	X	ASEQ	Inklusive Lokaldaten
eDOC	X	X	X		X	ASEQ	Inklusive Lokaldaten
Google Scholar	X	X	X	X		XML	Inklusive Anzahl der Bibliotheken und Link zum Verbundkatalog

eDOC: Lokaldaten werden für ‚Lokale Sicht‘ verwendet

Google: - Auflösung der MAB-Hierarchien über eDOC ‚MAB2-2-DC‘ Tool, dann XML-Konversion
- Google-definiertes Export-Format, mehrere Dateien mit ca. 1 MB

Registrierung von BIB- und HOL-Änderungen

Registrierung von Änderungen in bibliographischen und ZS-Bestandsdaten

Wie registriert Aleph Änderungen?

- **Benutzerdefiniert**

Fix_doc_mab_date

Bildet automatisch die MAB-Felder

- 002a Erfassungsdatum
- 003_ Änderungsdatum
- 070_ Produzierende Bibliothek
- 070b Ändernde Bibliothek

Probleme:

- Fix muss eingerichtet sein (problematisch bei Daten aus den Lokalsystemen - HOL-Daten)
- Datumsfelder können trotzdem manipuliert werden

- **Systemdefiniert**

Feld CAT

```
CAT |a BEARBEITER |b 20 |c 20060817 |1 ACC01 |h 0932
```

Lösung: Aufbau eines Direct-Index des CAT-Feldes – Subfeld c

CAT Indekskürzel für BIB-Library
CATH Indekskürzel für HOL-Library

Registrierung von BIB- und HOL-Löschungen

Registrierung von Löschungen in bibliographischen und Bestandsdaten

Was ist in Aleph ein gelöschter Satz?

Kennzeichen: **Kat. DEL, Kat. 001 ist entfernt**

Beispiel:

```
LDR 00809nM2.01200024-----h
FMT MH
DEL |a Y
CAT |a BATCH |b 20 |c 20060801 |l ACC01 |h 1904
SYS 000132031
```

Wie registriert Aleph Löschungen?

Oracle-Tabelle z00H

Problem:

- z00H wird nur bei Online-Löschungen beschrieben!
- Löschungen über p_manage_18 werden nicht registriert!
- Struktur von z00H analog z00 = sql-unfreundlich!

Lösung OBVSG: Einrichtung eines Oracle-Triggers auf den Directindex z11

Registrierung von Z300 Änderungen / Löschungen

Registrierung von Änderungen / Löschungen in der Exemplartabelle z300

KEINE Unterstützung von Aleph vorhanden!

Ausnahme:

p_file_13 - Job zur Extraktion der z300-Tabelle

Problem:

- NUR vollständige Extraktion der z300-Tabelle möglich – 10 Mio. Exemplare!!
- Ergebnis UNIX-File!

Lösung OBVSG:

- Einrichtung eines Oracle-Triggers zur Registrierung aller z300-Update-Vorgänge
- Neue / geänderte / gelöschte Sätze

Export der Lokaldaten

Herausforderung Lokaldaten

Obwohl jetzt alle Änderungen in den Lokaldaten (HOL / Z300) registriert werden, gibt es noch keinen einfachen Weg in Aleph, diese Daten zu extrahieren und exportieren!!

Grundsätzlich sind 2 Wege möglich

1. Extraktion der geänderten Daten – HOL und Z300
- komplex, gutes EDV-Wissen notwendig, Anwendung nicht effizienter Aleph-Tools - p_file_13
2. Anwendung der speziellen, vom hbz beauftragten Expand-Routine `expand_doc_bib_088_hbz` mit `Service p_print_03` welche Lokaldaten (HOL und Z300) in den BIB-Record expandiert – `MAB 088`

... wir verwenden die 2. Lösung!

Beispiel: p_print_03-Export mit hbz-Expand

```
000003478 001 L $$aAC00512306
000003478 331 L $$aControl theory, numerical methods and computer systems modelling
000003478 410 L $$aBerlin [u.a.]
000003478 412 L $$aSpringer
000003478 425a L $$a1975
000003478 540a L $$a0-387-07020-6
000003478 088 L $$aEFBM$$bm$$c202506 I.107$$eentlehnbar 4 Wo ←
000003478 088 L $$aBSH$$c35:B474$$eFreihandentl. ←
000003478 088 L $$aCFBN$$bF$$cN2737/107$$eentlehnbar,wenn kein Datum ←
000003478 088 L $$aKHB$$bMAG1$$c36859-B.107$$eentlehnbar MAG ←
```

Workflow Export Suchmaschinen - 1

Workflow Export Suchmaschinen - Dreiländerkatalog

1. Ermittlung der IDs aller relevanten BIB-Records als Input für p_print_03 (SQL-Abfrage)
Beispiel Datenexport 2006-05-06 – 2006-06-18

1a	BIB neu + upd	Index CAT	214.790
1b	HOL neu + upd	Index CATH	10.909
1c	HOL del	Oracle trigger	922
1d	z300 neu + upd + del	Oracle trigger	301.091
		Summe	<u>527.712</u>

- 1e Deduplizierung der IDs **457.854**
2. Export mit p_print_03
3. Ermittlung der IDs der gelöschten BIB-Sätze aus Oracle Trigger-Tabelle
4. p_print_03-Export UND Liste der gelöschten BIB-Sätze an hbz

ASEQ-Export-File: 589 MB

Run time: 3 Stunden

Workflow Export Suchmaschinen - 2

Workflow Export Suchmaschinen - eDOC

1. Ermittlung der IDs aller relevanten BIB-Records als Input für p_print_03 (SQL-Abfrage)
Analoge Schritte wie bei Dreiländerkatalog

1a	BIB	neu + upd	Index CAT
1b	HOL	neu + upd	Index CATH
1c	HOL	del	Oracle trigger
1d	z300	neu + upd + del	Oracle trigger

- 1e Deduplizierung der IDs

2. Die IDs von 1a bilden die Basis zur Erzeugung von ACC01.Z104 (trigger) Sätzen zum Update der eDOC Meta-Sätze
3. Die Daten von 1b,1c und 1d bilden die Basis zum Update der eDOC-Nutzungsdaten (Basis für die lokale eDOC Web-Sicht)
4. Zur Entfernung gelöschter BIB-Sätze aus eDOC wird ein eigenes Verfahren verwendet (zugehörige Objekt-Sätze müssen kontrolliert entfernt werden)

Workflow Export Suchmaschinen - 3

Workflow Export Suchmaschinen – Google Scholar

- Google unterstützt KEINE inkrementelle Updates
- Der volle Datenexport sollte monatlich durchgeführt werden

Ein erster Vollexport wurde im Juli 2006 durchgeführt

4.281 ~1MB Files repräsentieren 4.113.438.00 Verbunddatensätze (Run Time: 54 Std.)

1. p_print_03-Vollexport aller BIB-Records - NUR 2 Felder: 001, 088 (Run Time: 20 Std.)
 - zur weiteren Verarbeitung werden nur Sätze mit Lokaldaten herangezogen!!
2. Daten-Export aus dem Verbundkatalog in DC-format mit Auflösung der Hierarchien
 - eDOC-Routine vorhanden (Run Time: ca. 20. Std.)
3. Daten-Konversion in das Google-XML-Format - UTF-8, kleines Subset der BIB-Felder
 - jeder BIB-Satz muss die Anzahl der nutzenden Bibliotheken beinhalten - aus MAB 088
 - (Run Time: ca. 14. Std.)
4. Herstellung des "directory file" für die Google crawler - Google download Zeit = 5 Std.

Datenverfügbarkeit September 2006!

Zusammenfassung

Zusammenfassung

- Die Fähigkeit, Daten aus dem Bibliothekssystem zur Versorgung externer Anwendungen wie Suchmaschinen, externe Kataloge etc., zu exportieren, ist eine wichtige und mehr denn je benötigte Funktionalität
- Diese Aufgabe ist, wie wir zeigen wollten, nicht trivial und verlangt hohe Aleph- und EDV-Expertise sowie die entsprechenden Ressourcen
- Im Zuge der Primo-Entwicklung arbeitet Ex Libris unter anderem gemeinsam mit dem hbz auch an der Entwicklung von Daten-Extract-Mechanismen für Aleph-Kataloge

... darüber werden wir im nächsten Vortrag mehr erfahren!