CC-interop: A Post Mortem

http://ccinterop.cdlr.strath.ac.uk/

George Macgregor & Gordon Dunsire, 
Centre for Digital Library Research, 
Department of Computer & Information Sciences
Introduction

• George: Introduction / Background and Work Package A

• Gordon: Work Package B (Work Package C) and the Future!!!!

All project reports are available at:
http://ccinterop.cdlr.strath.ac.uk/
What is a Distributed Catalogue?

• Each institution has a database located at the heart of their Library Management System
• This database can be accessed from outside the institution
• Searches can be performed using Z39.50:
  - ‘Z’: information retrieval protocol
  - A ‘broadcast search’ can be conducted (involves searching multiple databases / targets simultaneously)
• Software gathers results from the remote databases and presents them to the user
• Search can be a sub-set of databases available (e.g. CAIRNS has 1-19+, InforM25 1-36 in and around London area)
What is COPAC?

- **COPAC** = the CURL OPAC
- Institutional databases copied and ‘fused together’
- Thus producing a single, mammoth, database
- Weekly data loads
- 26 UK library members, including BL
- Administered by MIMAS, Manchester Computing
- On behalf of the JISC
CC-interop: A Post Mortem - Strathclyde University, 13th October 2004

Remote databases in library systems

Clump
Software & User interface

Internet

COPAC
Single, large database

USER

Internet
CC-interop Project

- CC-interop = “COPAC/Clumps Continuing Technical Cooperation Project”
- Funded by JISC via the JISC Committee for the Information Environment
- Duration: May 2002 - April 2004 (Final Report Submitted to JISC in July 2004)
- Three work packages:
  - WP A - M25 Systems Team & MIMAS
  - WP B - CDLR [& RIDING]
  - WP C - CERLIM & project partners
WP A

• Thorough technical investigations of cross-searching/linking between different architectures

• Tasks:
  - Comparing how searches are carried out at target database
  - Analysis of record retrieval process
  - Performance testing
  - Detailed technical analysis of ‘combined’ architecture options
WP B

Using CAIRNS (CDLR) and RIDING clumps with the SCONE Collection Level Description (CLD) service for:

- Investigating and specifying collection description standards requirements
- Looking at CLD schemas in relation to both the clumps and COPAC
- Looking at the intelligent selection of databases in clumps by CLDs, based on dynamic landscaping
- Working towards guidelines for coping with variations in cataloguing & indexing practices to facilitate interoperability [between the clumps and COPAC]
WP C

• User Behaviour Study - area such as:
  - What do users do when they search large union catalogues?
  - Do they understand what it is they are searching?
  - Do they find what they are looking for?
  - What features would they like to see?

• CERLIM (MMU)
  - 1:1 user sessions at 3 partner sites
    - Pre-search questionnaire
    - Recorded searches of ‘local’ clump and COPAC (Snag It)
    - Interview immediately after to discuss their experience
  - 3 focus groups of librarians
    - Set of 10 questions about a range of issues

• Report available on the project web site!
But, to what end?

• To continue work undertaken by previous JISC funded programmes, eLib Phase 3, etc. Component of the Research Libraries Network (RLN)
• UK National Catalogue (formerly known as UKNUC):
  - Still on the JISC agenda
  - Likely to incorporate national, university and large public libraries
  - Likely to be a mix of physical and distributed architectures
• To complement the Serials Union Catalogue
  - SUNCAT project at EDINA
WP A

As mentioned, the primary remit of WPA was to investigate interoperability between union catalogues of distributed and non-distributed architectures.

This entailed:

• Investigating whether both models could be connected (i.e. adding a clump to COPAC and vice versa)

• Investigating relevant issues pertaining to searching performance, results issues, landscaping, etc.
WP A Method

- InforM25 Copy (CC25): added as COPAC Z-target

- Deployment of JAFER as middleware: ‘Java Access to Electronic Resources’ developed at Oxford for JISC 5/99
  
  • Free Open Source software
  
  • Customised for the purposes of CC-interop (Logging facilities augmented, Extensible Stylesheet Language Transformations (XSLT), Concatenations (mini-clump))
WP A Method (cont.)

COPAC Interface Copy: Enable independent logging, etc.

Results & Display Issues: Detailed analysis of COPAC search result manipulation and display issues. Could they be applied in a distributed environment?
Outputs & Results (WP A)

Semantic interoperability & index composition

• Technical interoperability relatively ‘easy’, but limited semantic interoperability

• Disparate cataloguing & indexing practices impairing semantic interoperability (detailed findings & analysis of conclusions outlined in the CAIRNS final report)

• COPAC exploits features peculiar to physical union models (COPAC can enrich indexing, thus potentially improving the retrieval of relevant records)
Outputs & Results (WP A) (cont.)

Technical interoperability

- JAFER meets many of the needs for distributed catalogue services & could be used by the clumps. Further exploitation of JAFER recommended in IE. *(JAFER further investigated by CREE (Contextual Resource Evaluation Environment) as we speak…)*

- Technically possible to landscape using JAFER as middleware

- Query reconfiguration can be carried out within the middleware to ensure optimal searching of different Z-targets *(although this functionality would not be necessary if there was wider adoption of the Bath Profile)*
Outputs & Results (WP A) (cont.)

Results processing

• Problems with record matching, de-duplication, consolidation, ranking in most distributed services

• COPAC on-the-fly routines could feasibly be applied to the clumps (such routines would possibly benefit from revision to reflect rapidly changing user behaviour – see WPC, work of CIBER)
  • Further testing is needed as the algorithms developed by COPAC would add value to results display
  • Transaction time: Is a trade off is needed?
Response Times

• 90% of response were received in under 1 second, with some responding in less than 0.125 seconds; Broad & fast times worthy of further investigation
• **No servers** showed slower response times during what would be consider ‘peak’ periods of heavy use of the local OPAC (i.e. mid-morning to early evening)
• Generally good performance: response problems the result of non-response and how this is handled by the client software

**Further investigation**: short time-outs & **MORE** user research; response times & Boolean; quick & dirty Z installations;
Over to Gordon.....