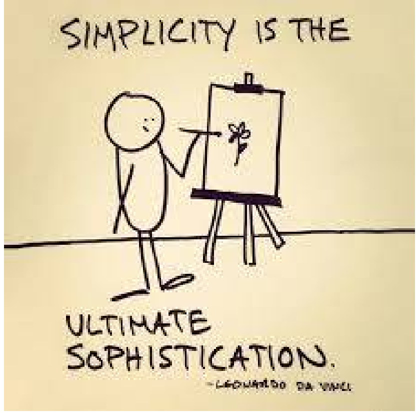


Investigating the effectiveness of ***client-side search/browse*** *without a network connection*

Hussein Suleman <hussein@cs.uct.ac.za>
Digital Libraries Lab, Department of Computer Science, School of IT
University of Cape Town

6 November 2019

Repositories in Low-Resource Environments

- What if you want to create an archive to store heritage collections, with typical discovery services,

  BUT

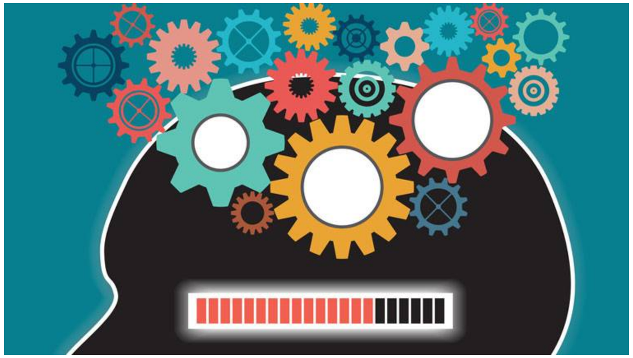
  you are in a **low-resource environment**?

What is a Low-Resource Environment?

- Poor countries
  - + Poor regions in rich countries
    - + some scenarios in rich regions in rich countries

- As examples:
  - Malawi
    - Rural Scotland
      - Someone in New York City wanting to curate family photos

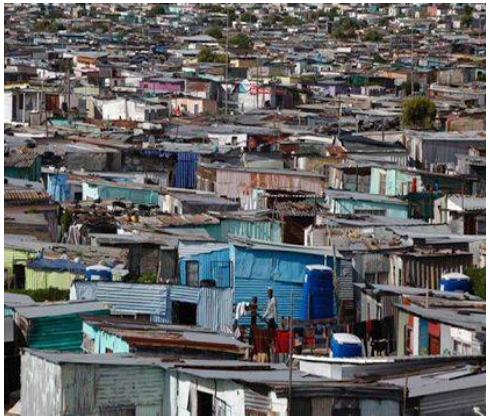
Low Resource Countries 1/3



Skills and Education

- Typical archivists are not as highly skilled as counterparts elsewhere.
- Digital media is still not the norm.
- Education levels of general population hinders preservation – end-user data curation is very difficult.

Low Resource Countries 2/3



Funding

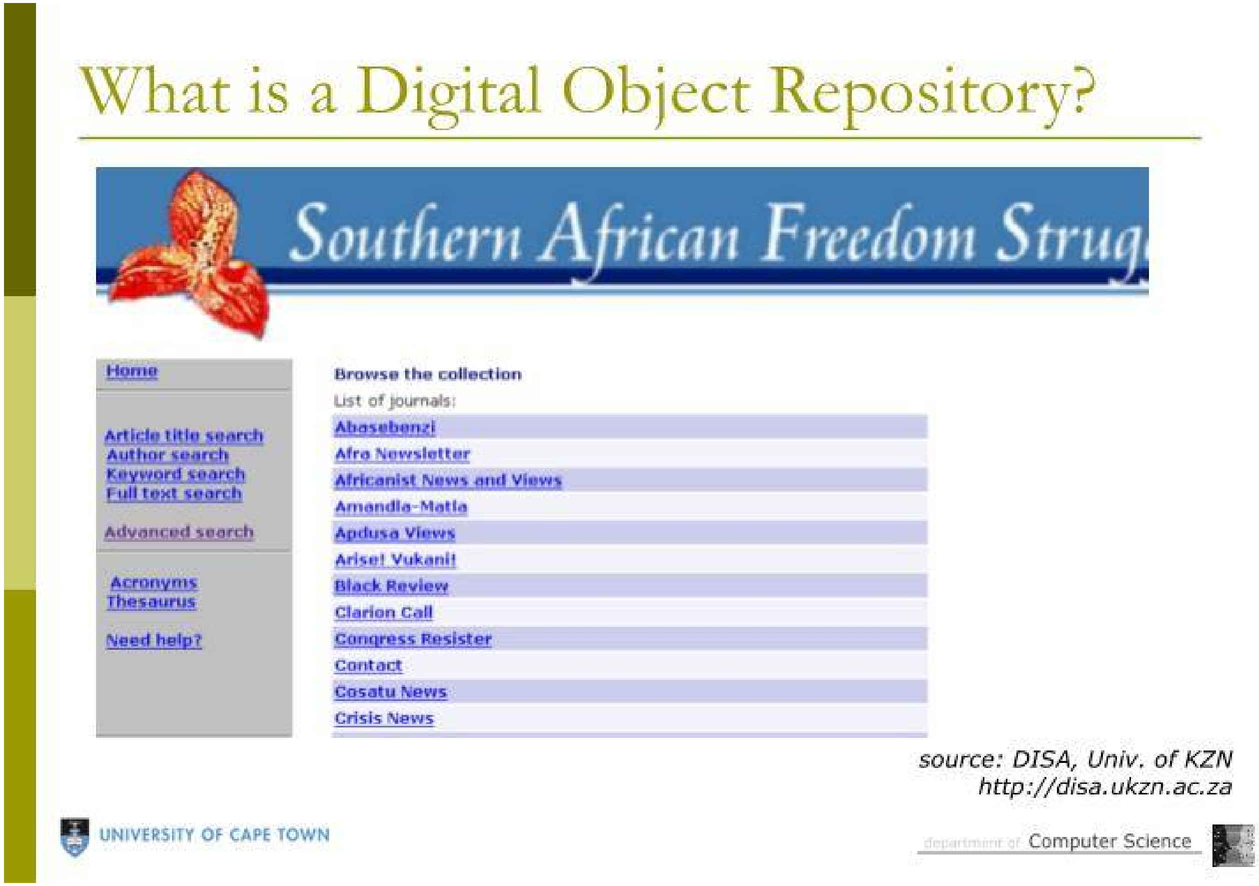
- Typically, there is little.
  - If we had money, there are other priorities ...
- Many preservation projects are funded by external agencies, but with restrictions on data accessibility.
- There is a desperate need to do more with less.

Low Resource Countries 3/3

Internet Bandwidth

- Non-existent in some places and not as good everywhere else.
- Preservation projects designed for high bandwidth are not suitable.
- All online solutions must be bandwidth-friendly.

What is the net effect? 1/3



What is a Digital Object Repository?

Southern African Freedom Strug

Home

Article title search
Author search
Keyword search
Full text search

Advanced search

Acronyms
Thesaurus

Need help?

Browse the collection

List of journals:

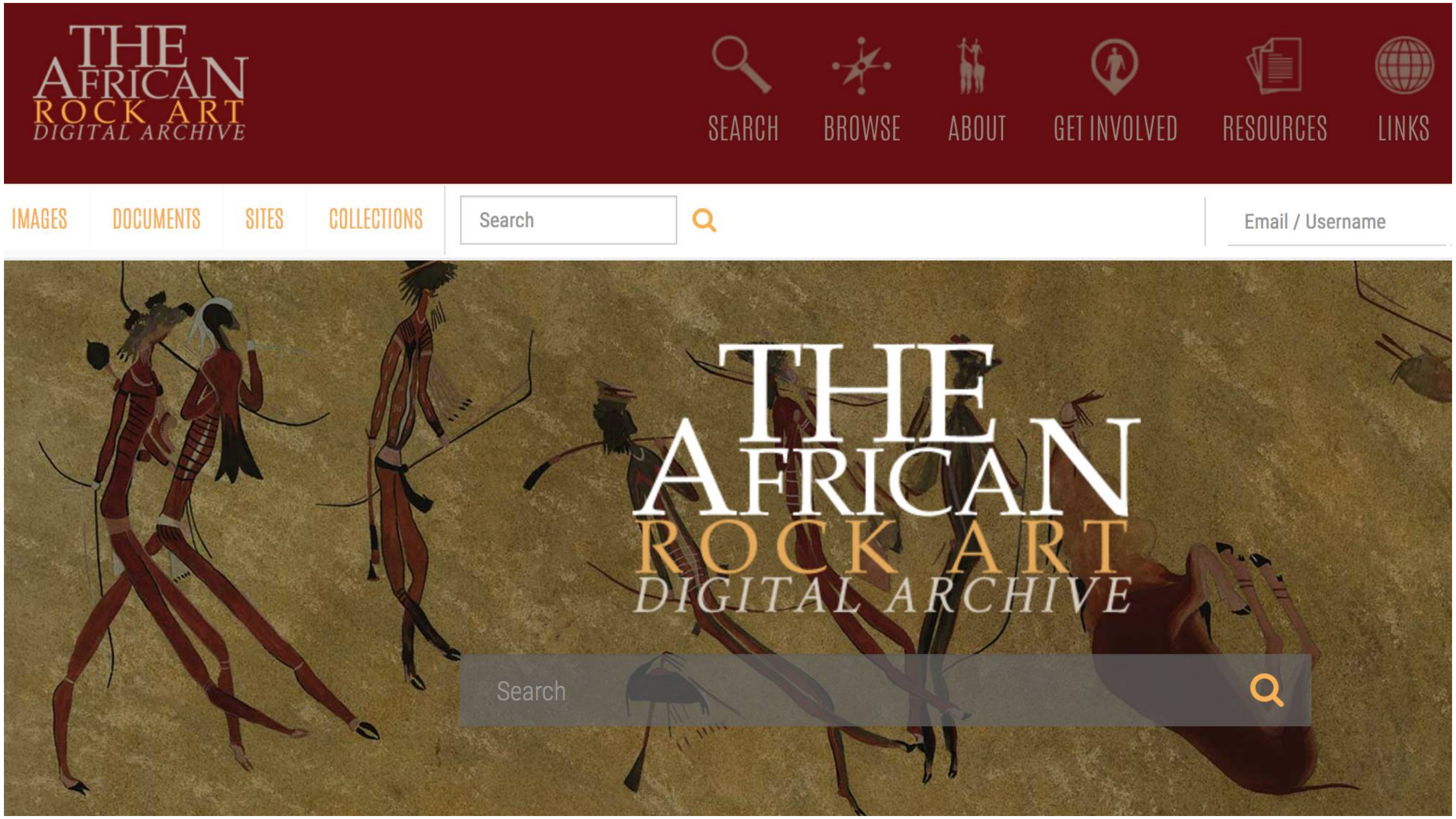
- Abasebenzi
- Afro Newsletter
- Africanist News and Views
- Amandla-Matla
- Apdusa Views
- Arise! Vukani!
- Black Review
- Clarion Call
- Congress Resister
- Contact
- Cosatu News
- Crisis News

source: DISA, Univ. of KZN
http://disa.ukzn.ac.za

UNIVERSITY OF CAPE TOWN

department of Computer Science

What is the net effect? 2/3



What is the net effect? 3/3

How to build a Low Resource Archive

- *"2 million euros and 2 years and we can build any digital repository system"*
- Can we use DSpace/AtoM/etc.?

- Can we do the same thing as everyone else?

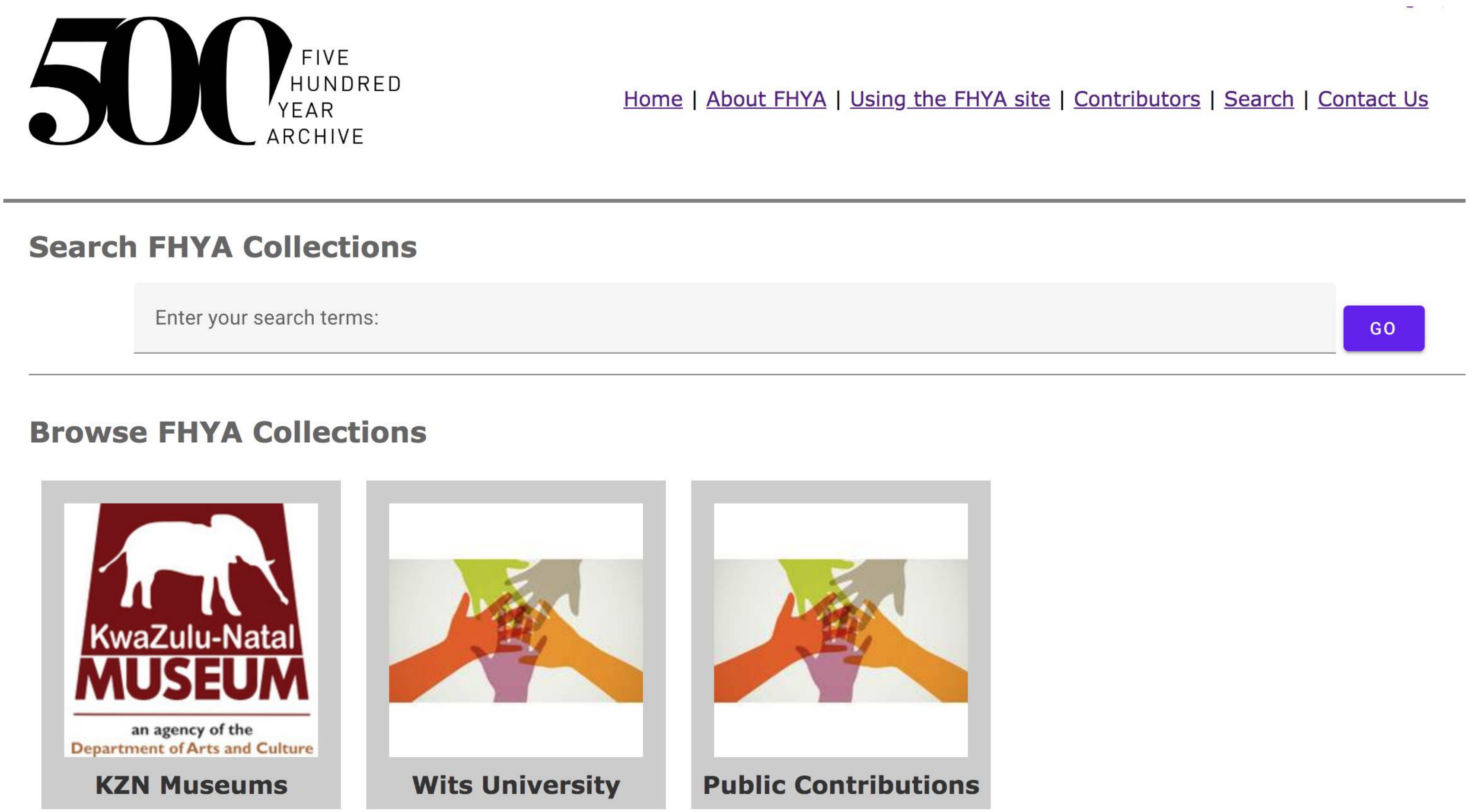
  OR

- Can we create a more suitable architecture for low resource environments?

Surely someone did this already ...

- **Greenstone**
  - Distributable on a CDROM; installs system locally.
- **Project Gutenberg**
  - Philosophy that we keep things simple and they will last.
- **Bleek and Lloyd Collection**
  - Simple Javascript search of website.

FHYA Prototype 1/3



Home | About FHYA | Using the FHYA site | Contributors | Search | Contact Us

Search FHYA Collections

Enter your search terms:

GO

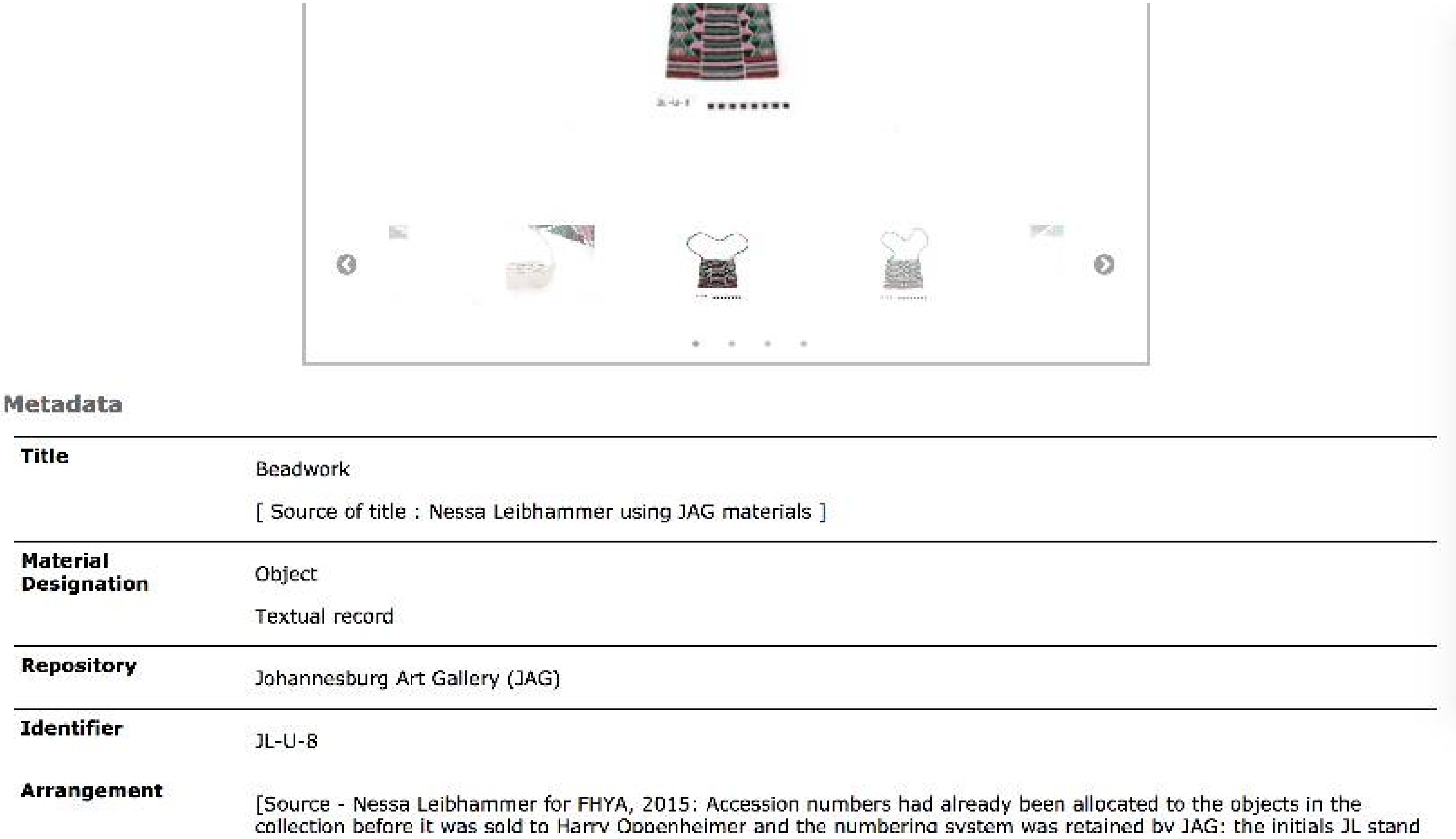
Browse FHYA Collections

KZN Museums

Wits University

Public Contributions

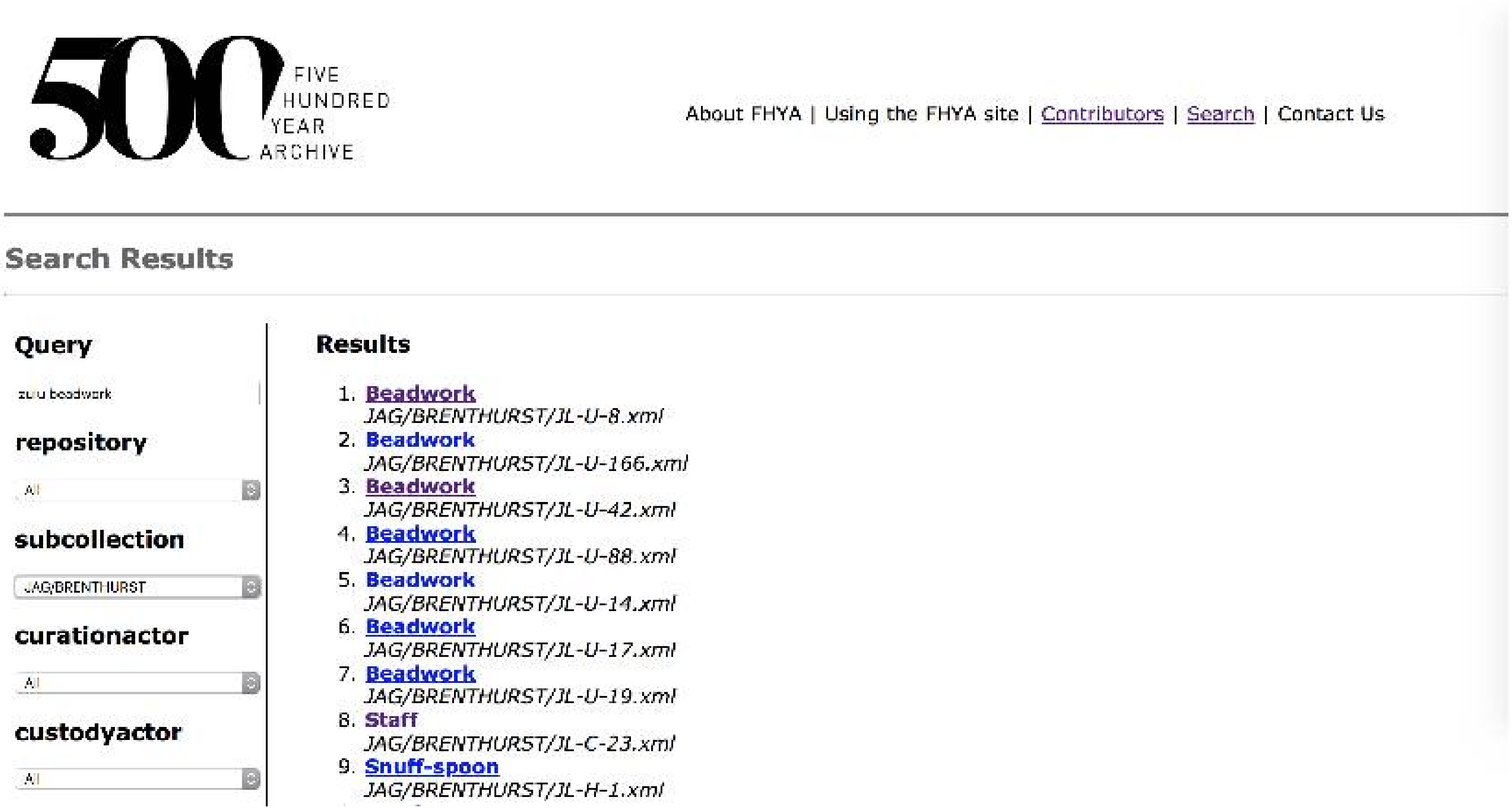
FHYA Prototype 2/3



Metadata

Title	Beadwork [Source of title : Nessa Leibhammer using JAG materials]
Material Designation	Object Textual record
Repository	Johannesburg Art Gallery (JAG)
Identifier	JL-U-8
Arrangement	[Source - Nessa Leibhammer for FHYA, 2015: Accession numbers had already been allocated to the objects in the collection before it was sold to Harry Oppenheimer and the numbering system was retained by JAG: the initials JL stand

FHYA Prototype 3/3


500 FIVE HUNDRED YEAR ARCHIVE
About FHYA | Using the FHYA site | Contributors | Search | Contact Us

Search Results

Query

zulu beadwork

repository

All

subcollection

JAG/BRENTHURST

curationactor

All

custodyactor

All

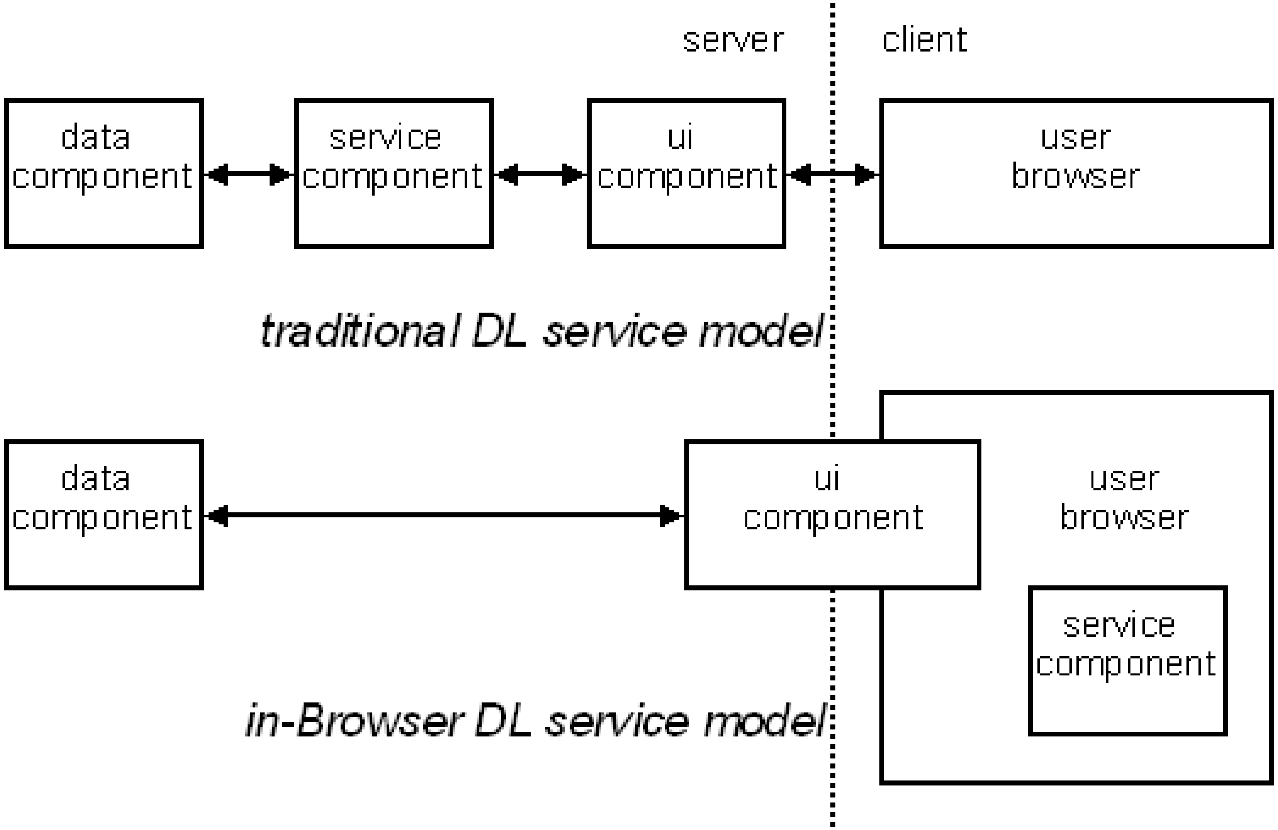
Results

1. Beadwork
   JAG/BRENTHURST/JL-U-8.xml
2. Beadwork
   JAG/BRENTHURST/JL-U-166.xml
3. Beadwork
   JAG/BRENTHURST/JL-U-42.xml
4. Beadwork
   JAG/BRENTHURST/JL-U-88.xml
5. Beadwork
   JAG/BRENTHURST/JL-U-14.xml
6. Beadwork
   JAG/BRENTHURST/JL-U-17.xml
7. Beadwork
   JAG/BRENTHURST/JL-U-19.xml
8. Staff
   JAG/BRENTHURST/JL-C-23.xml
9. Snuff-spoon
   JAG/BRENTHURST/JL-H-1.xml

Low Resource Repository ideas ...

- Minimalism
- Off-line or online access
- Data in structured files (e.g., XML), not DBMS
- Pre-generated interfaces where possible
- Preservation by copying
- OS independence
- Services on the client-side where possible

The in-Browser Service Model



How client-side search/browse works

- Step 1:
  - All metadata stored in files.
- Step 2:
  - Indices created and stored in files.
- Step 3:
  - Query processing in Javascript.
  - UI partly pre-generated and partly updated using Javascript.

Notice – this will even work if you are offline!

Search/Browse Implementation Details

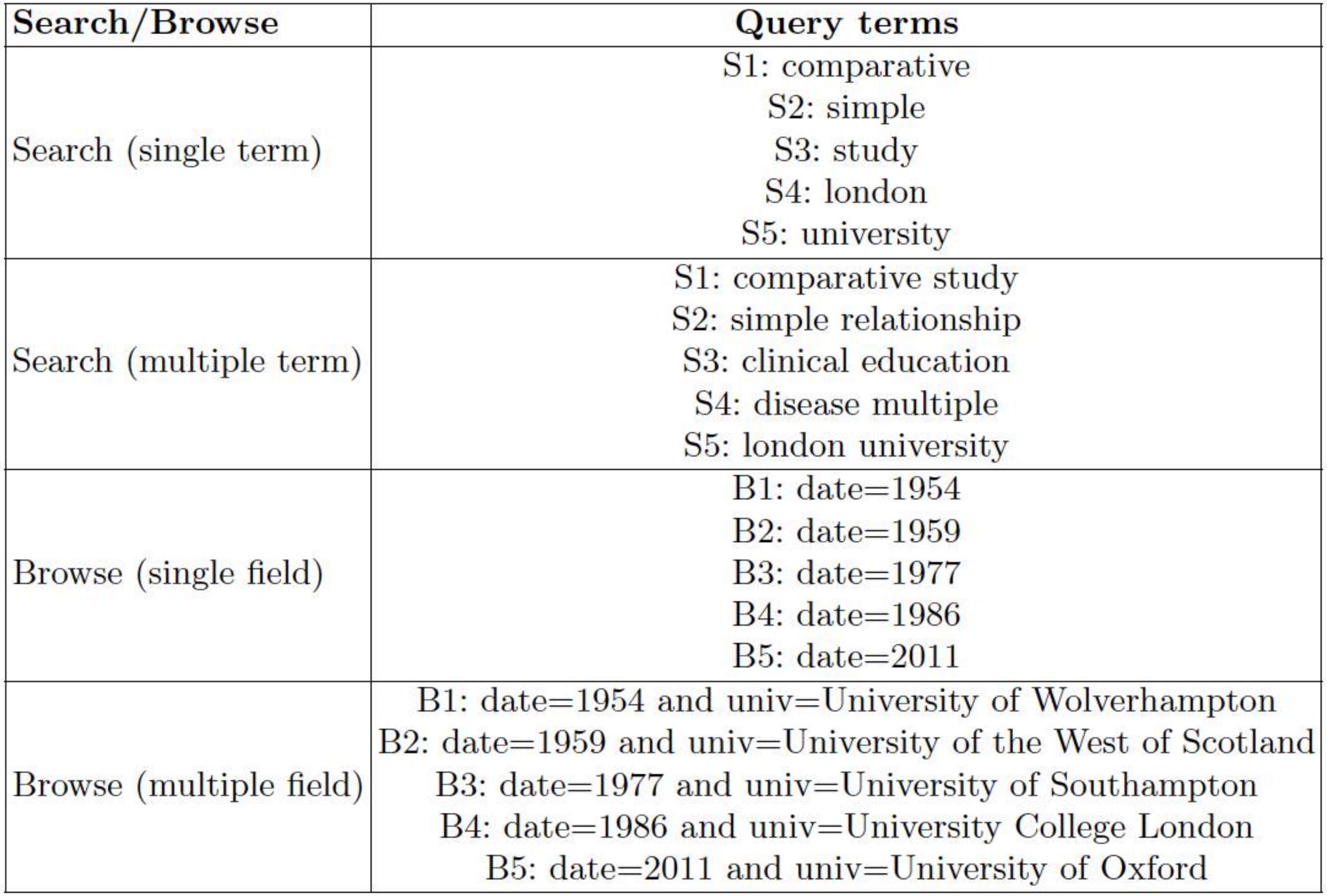
- Extended boolean search model
- + Faceted search
- Multi-term fielded query terms
  - "title:offline author:hussein"
- Stopwords, normalisation
- Configurable fields for search/facets
- Drop-down boxes for facets
- Multiple indices for different metadata subsets

Performance Experiment

- How well will it work? Surely browsers are too slow and collections too large? Hussein, this only works in really trivial cases!

- ETD metadata harvested from NDLTD.
- Test different collection sizes: 2000-32000.
- Test different typical operations, varying complexity.
  - Search, browse, search+browse

Queries



Search/Browse	Query terms
Search (single term)	S1: comparative S2: simple S3: study S4: london S5: university
Search (multiple term)	S1: comparative study S2: simple relationship S3: clinical education S4: disease multiple S5: london university
Browse (single field)	B1: date=1954 B2: date=1959 B3: date=1977 B4: date=1986 B5: date=2011
Browse (multiple field)	B1: date=1954 and univ=University of Wolverhampton B2: date=1959 and univ=University of the West of Scotland B3: date=1977 and univ=University of Southampton B4: date=1986 and univ=University College London B5: date=2011 and univ=University of Oxford

Index creation time

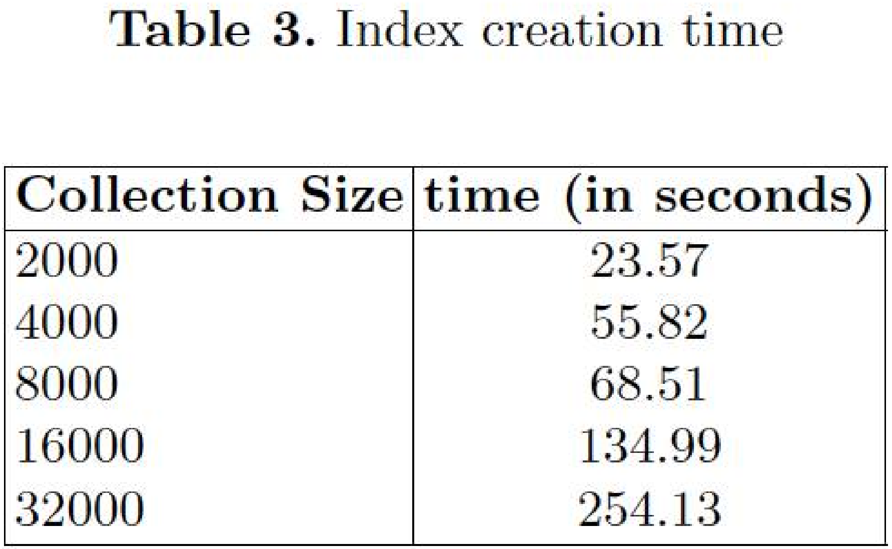


Table 3. Index creation time

Collection Size	time (in seconds)
2000	23.57
4000	55.82
8000	68.51
16000	134.99
32000	254.13

Simple Query Results – 16000 items

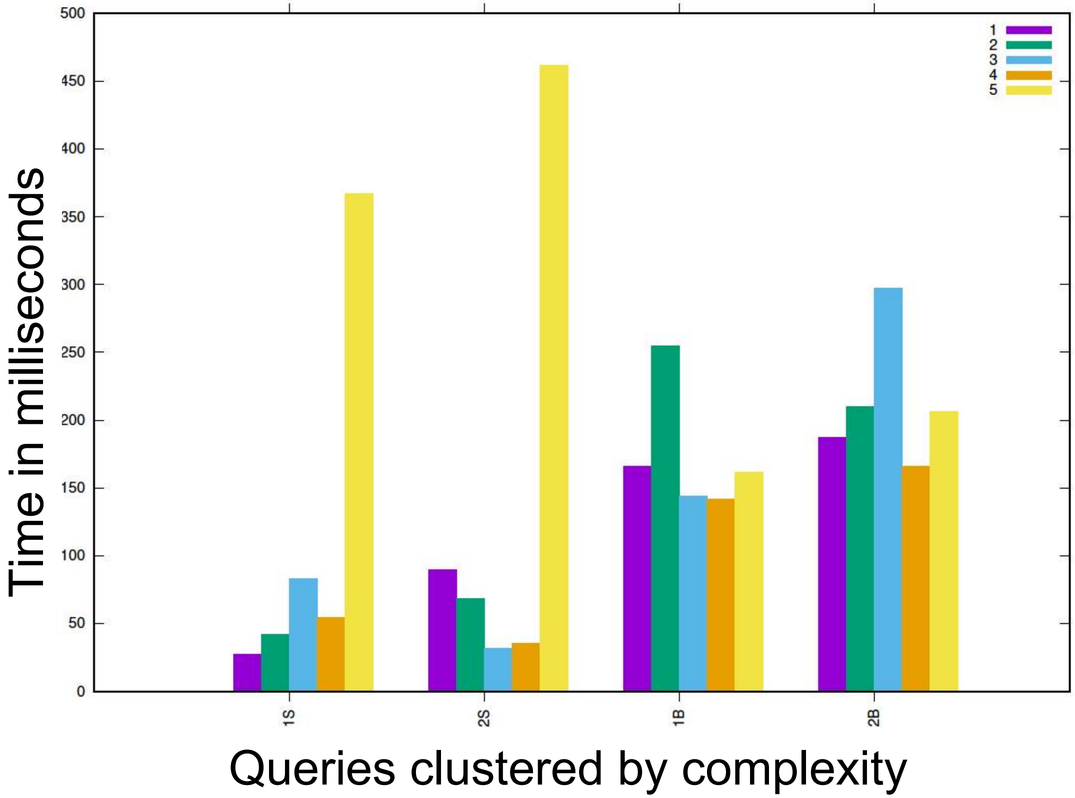


Fig. 2. Average times for queries of difference complexities. 1S/2S/1B/2B are the query complexities while the 5 data points within each cluster are the different queries tested.

Complex Query Results – Ave. Performance

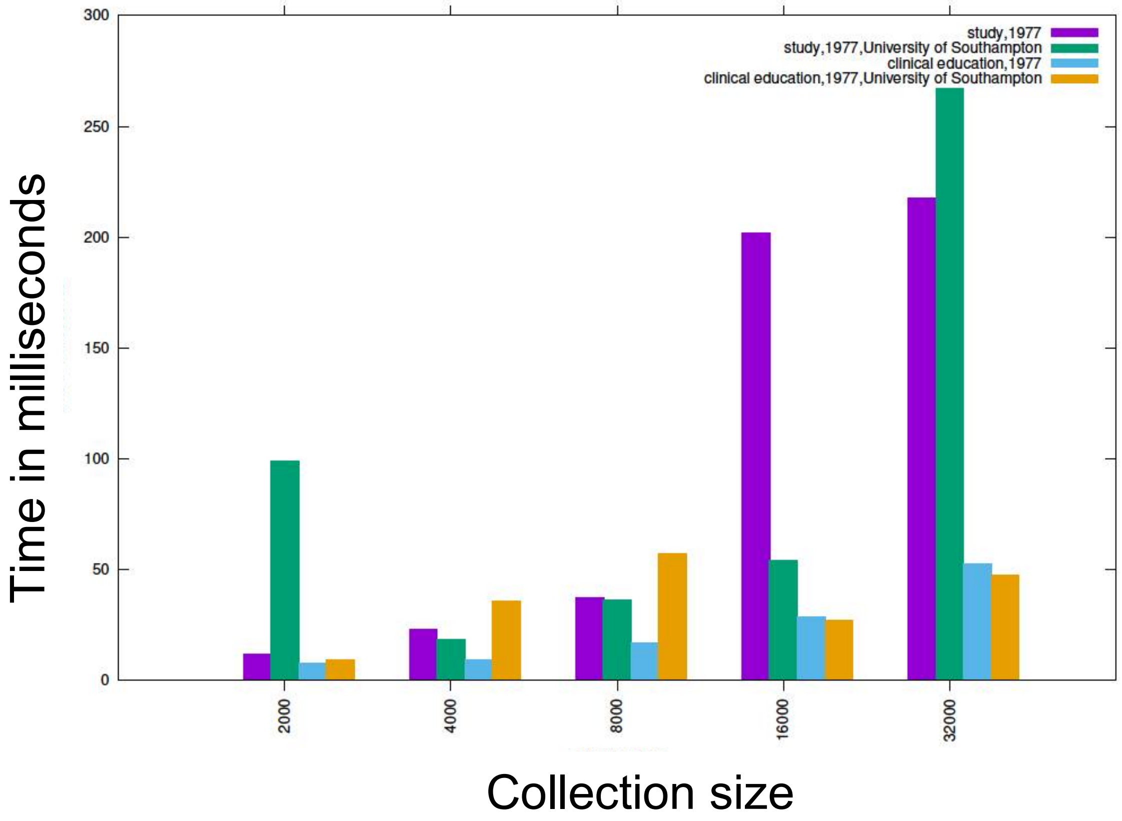


Fig. 5. Average times for faceted search/browse queries of difference complexities across all collection sizes.

Conclusions

- Performance scales with processing needed, with full indices/inverted files.
  - We could pre-compute more for better performance, but remember balance …
- In most cases, sub-second responses are possible for far more than 32000 records!
- There is little reason for complex DBMSes, server-side search engines, etc. for small collections.

Reflection

- One size does not fit all.
- Simple solutions for small problems.
  - Complex solutions for big problems (which aren’t as common as we think).
- Some ideas may lead to better preservation.
  - Only time will really tell…

- What if we built digital library systems differently?
  - Could we change those parts of the world that are still waiting?

that's all folks!