Flexible Design for Simple Digital Library Tools and Services

Lighton Phiri Hussein Suleman

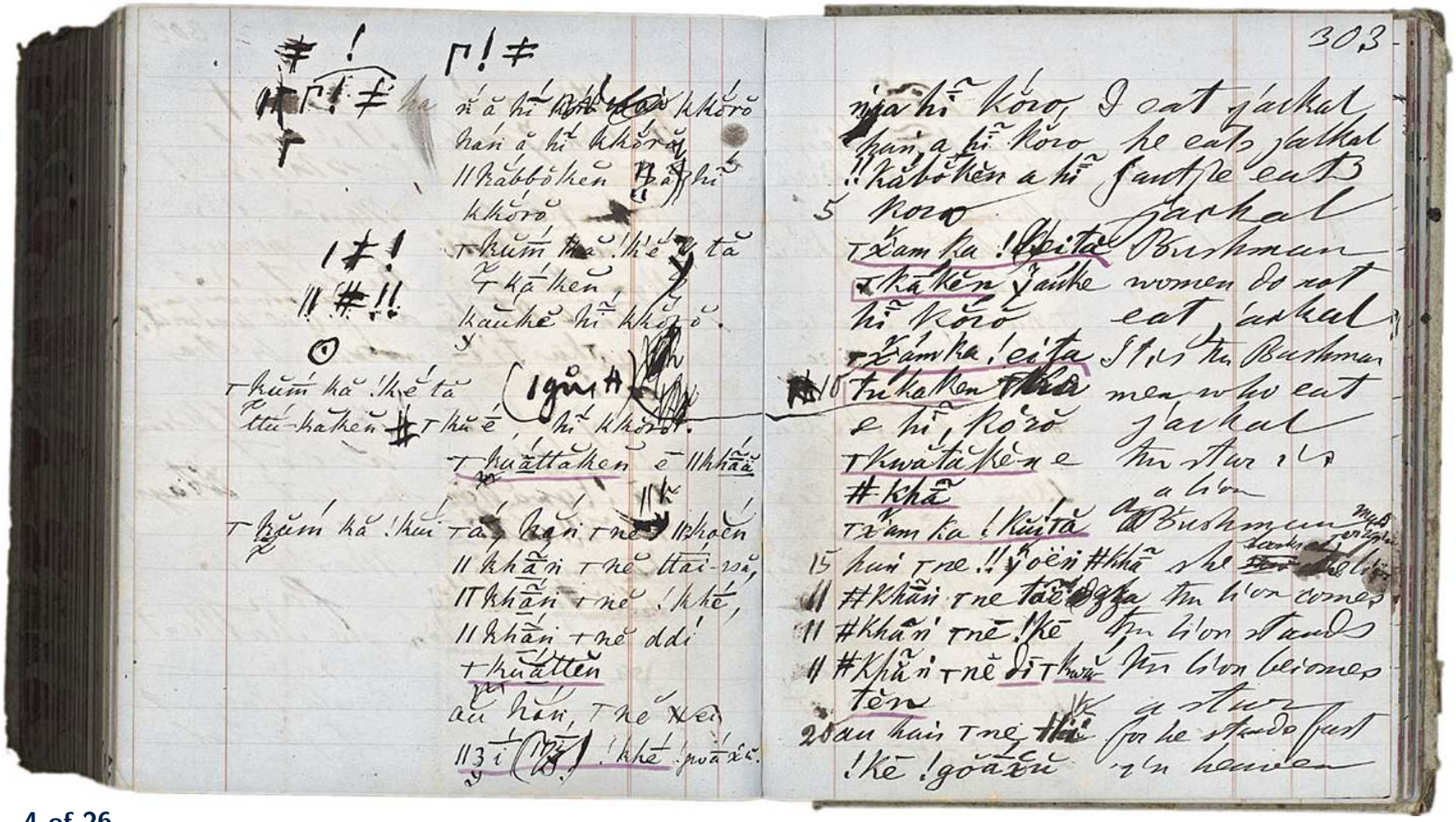
Digital Libraries Laboratory
Department of Computer Science
University of Cape Town

October 8, 2013

SARU archaeological database@ UCT

http://www.martinwest.uct.ac.za

http://lloydbleekcollection.cs.uct.ac.za



Contextual overview

- Problem and motivation
  - Preservation costs
  - Technical skills and education
  - Computing resources
- Proposed solution
  - Simplicity and minimalism
    - Successes of minimalism —Project Gutenburg
  - Principled DL design
- Prior work
  - Derivation of design principles
  - Repository implementation
    - Real-world case studies

Repository prototype architectural design

- File-based
  - Digital objects stored on native operating system
  - Hierarchical collection structure
- Metadata objects
  - Plain text files
  - Encoded using Dublin Core
  - Relationships modelled using metadata elements
- Object organisation
  - Metadata records stored alongside objects
  - Content objects and container objects nested within other container objects

User study experiment

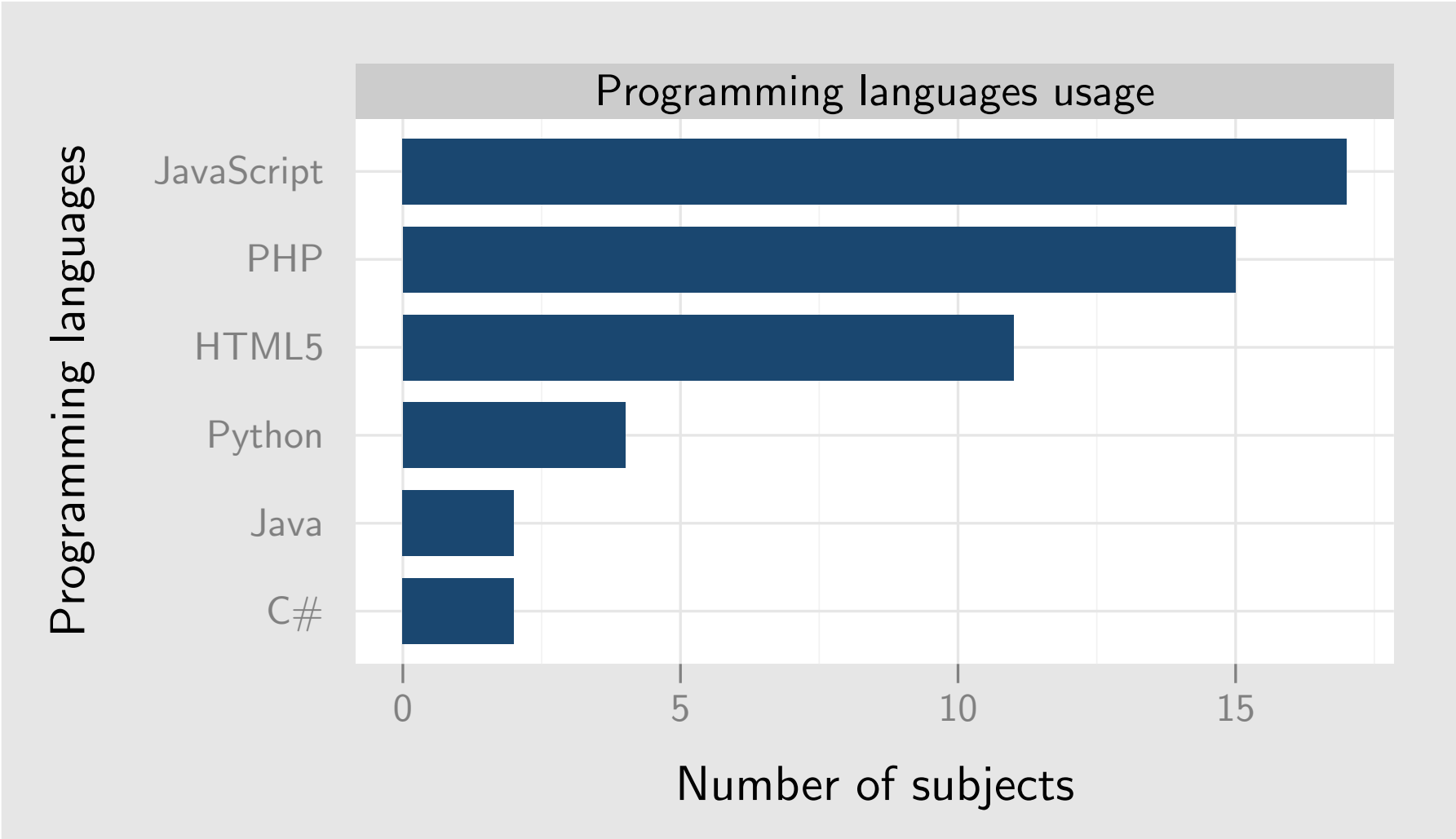
- Objective
  - Developer-oriented
  - Simplicity and flexibility of file-based store
- Target population
  - 34 Computer Science honours students
  - 12 groups of twos and threes
  - Skillset
    - Technologies relevant to study —DBMS, XML, Web apps
    - Storage solutions
    - Digital Libraries concepts
- Approach
  - Subjects tasked to build layered services using file-based store
  - Marks awarded for innovation —among other facets
  - Subjects answered post-experiment survey

User study experiment - results

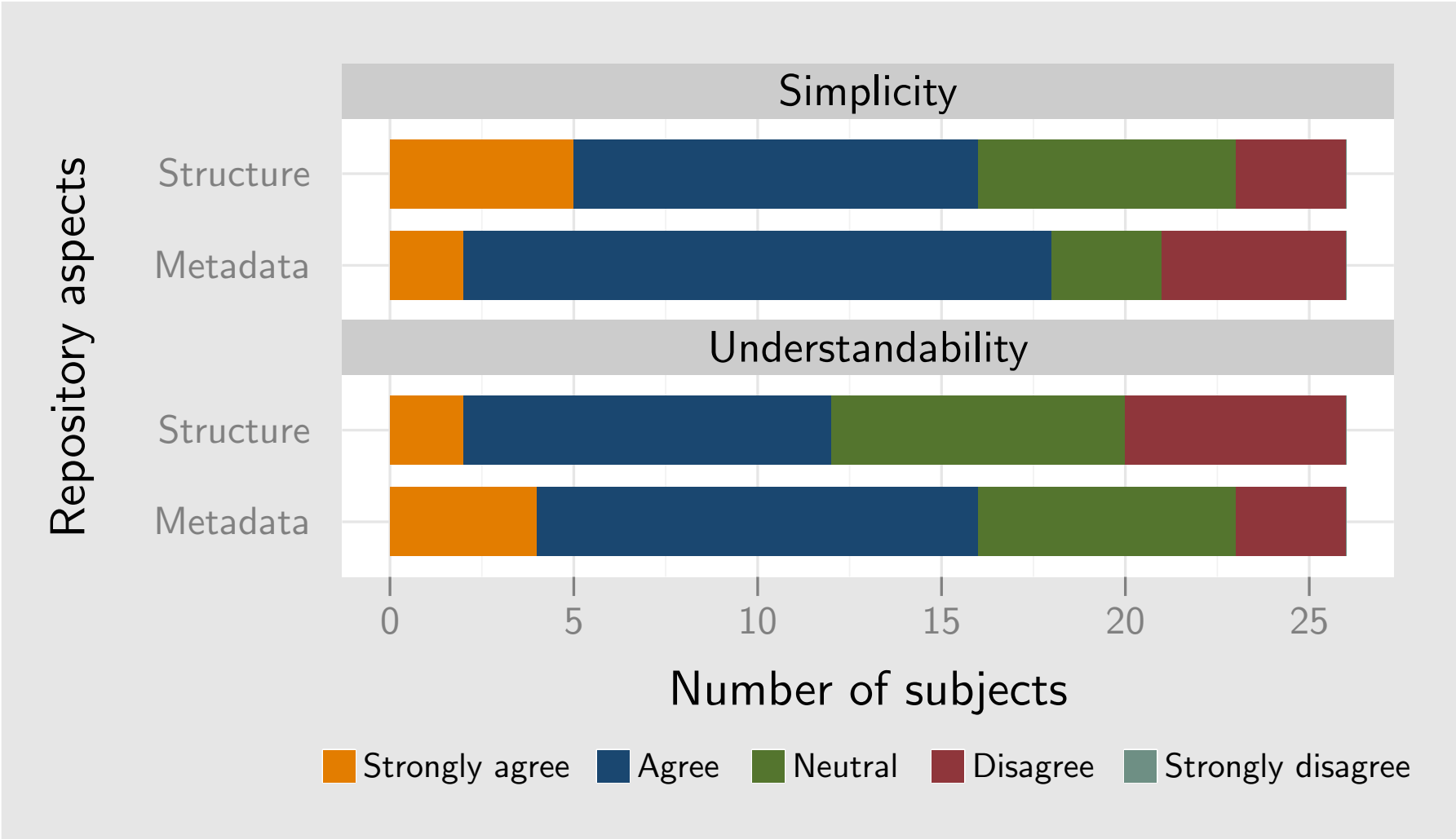
- Survey participants
  - 76% response rate —representation from all 12 groups

Group	Web service	Candidates	Respondents
Group 1	Transcription	3	3
Group 2	Downloader	3	3
Group 3	Commenting	3	1
Group 4	Visualisation	3	2
Group 5	Transcription	3	2
Group 6	Annotation	2	2
Group 7	Visualisation	3	3
Group 8	Browsing	3	3
Group 9	Annotation	3	2
Group 10	Rating	3	2
Group 11	Gestures	3	1
Group 12	Visualisation	2	2

User study experiment - results (1)



User study experiment - results (2)



User study experiment - results (3)

- Users asked to rank storage solutions in order of preference
- What aspects of your most preferred solution [database] above do you find particularly valuable?
  - "I understand databases better."
  - "Simple to set up and sheer control"
  - "Easy setup and connection to MySQL database"
  - "Ease of data manipulation and relations"
  - "Centralised management, ease of design, availability of support/literature"
  - "The existing infrastructure for storing and retrieving data"

User study experiment - results (4)

- Do you have any general comments about the data structure or format?
  - "Had some difficulty working the metadata, despite looking at how to process DC metadata online, it slowed us down considerably."
  - "Good structure although confusing that each page has no metadata of its own(only the story)."
  - "The hierarchy was not intuitive therefore took a while to understand however having crossed that hurdle was fairly easy to process."
  - "I guess it was OK but took some getting used to"

User study experiment - findings

- Simplicity resulted in more understandable structure
  - 69% agreed that XML-files were simple
  - 61% found XML format easy to work with
  - 62% found hierarchical structure simple to work with
  - 46% found hierarchical structure easily understandable
- Simplicity does not affect flexibility of interaction with file-store
  - No influence on choice of language
  - Only 15% of subjects thought it did

Performance experiment

- Objective
  - Assess performance relative to collection size
- Test Environment
  - Pentium(R) Dual-Core CPU E5200@ 2.50GHz; 4GB RAM
  - 32 bit Ubuntu 12.01 LTS
  - Siege and ApacheBench for benchmarking
- Metrics
  - Response time
- Factors
  - Collection hierarchical structure
  - Collection size —digital objects

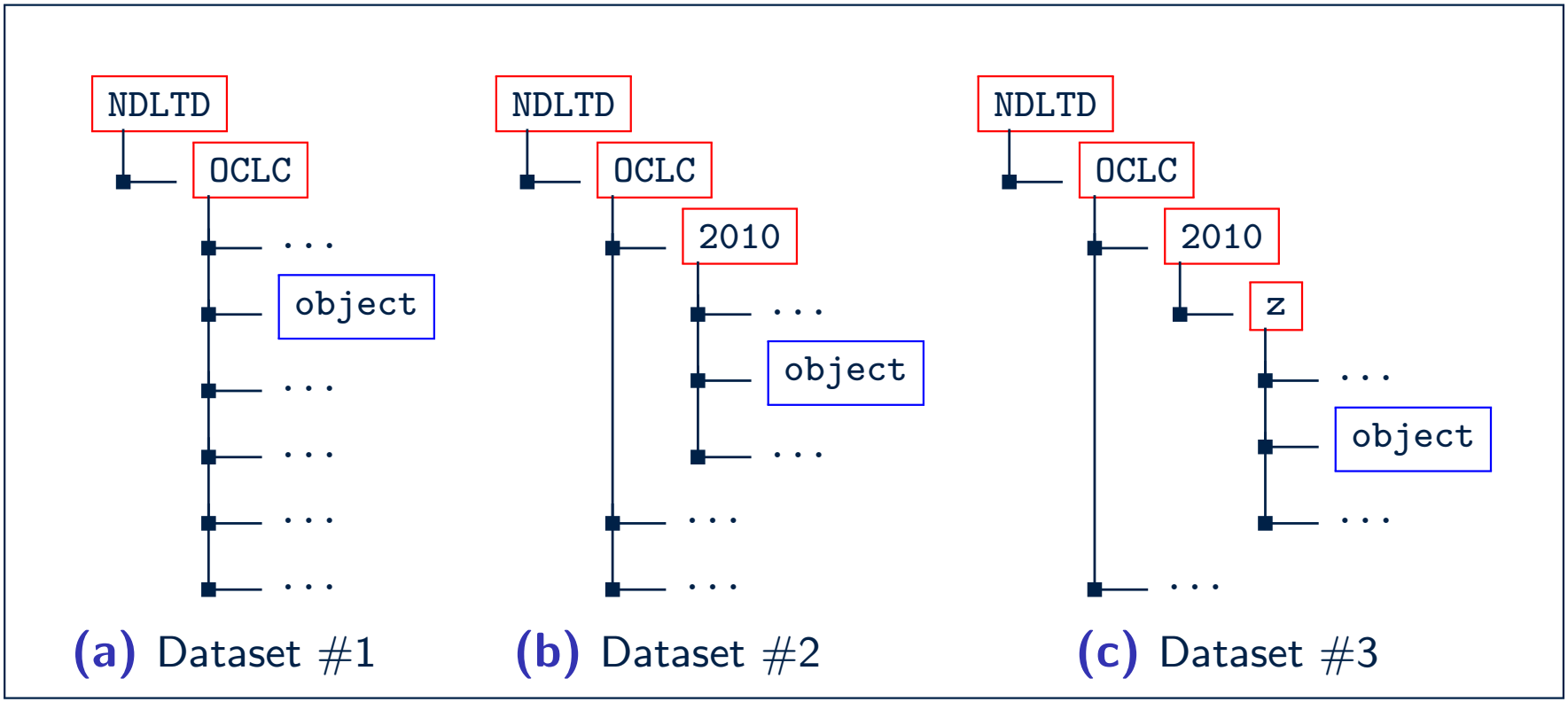
Performance experiment - test dataset

- NDLTD Union Catalog —`http://union.ndltd.org/OAI-PMH`
  - Harvested 1 907 000 metadata records
  - Dublin Core-encoded plain text files
- Linearly increasing workload

Workload	Objects	Cols	Size [MB]
W1	100	19	0.54
W2	200	25	1.00
W3	400	42	2.00
⋮	⋮	⋮	⋮
W13	409 600	128	1945.00
W14	819 200	131	3788.80
W15	1 638 400	131	7680.00

Performance experiment - test dataset (2)

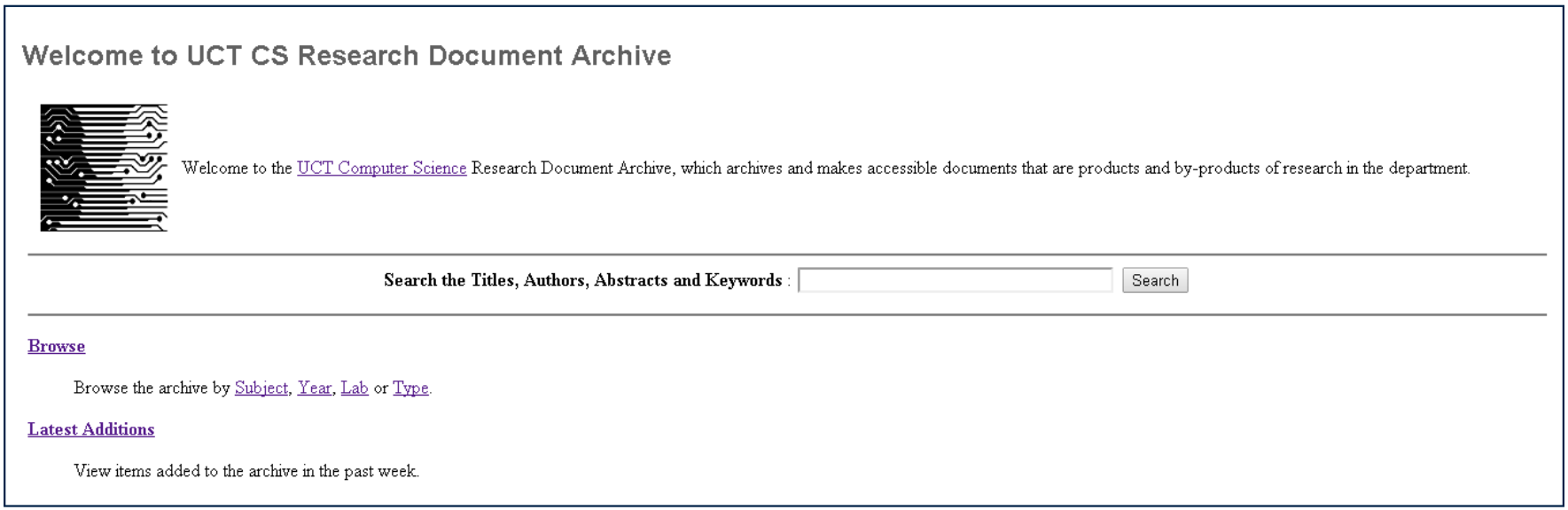
- Two datasets spawned from initial dataset
  - one-, two- and three-level structures



(a) Dataset #1 **(b)** Dataset #2 **(c)** Dataset #3

Performance experiment - evaluation aspects

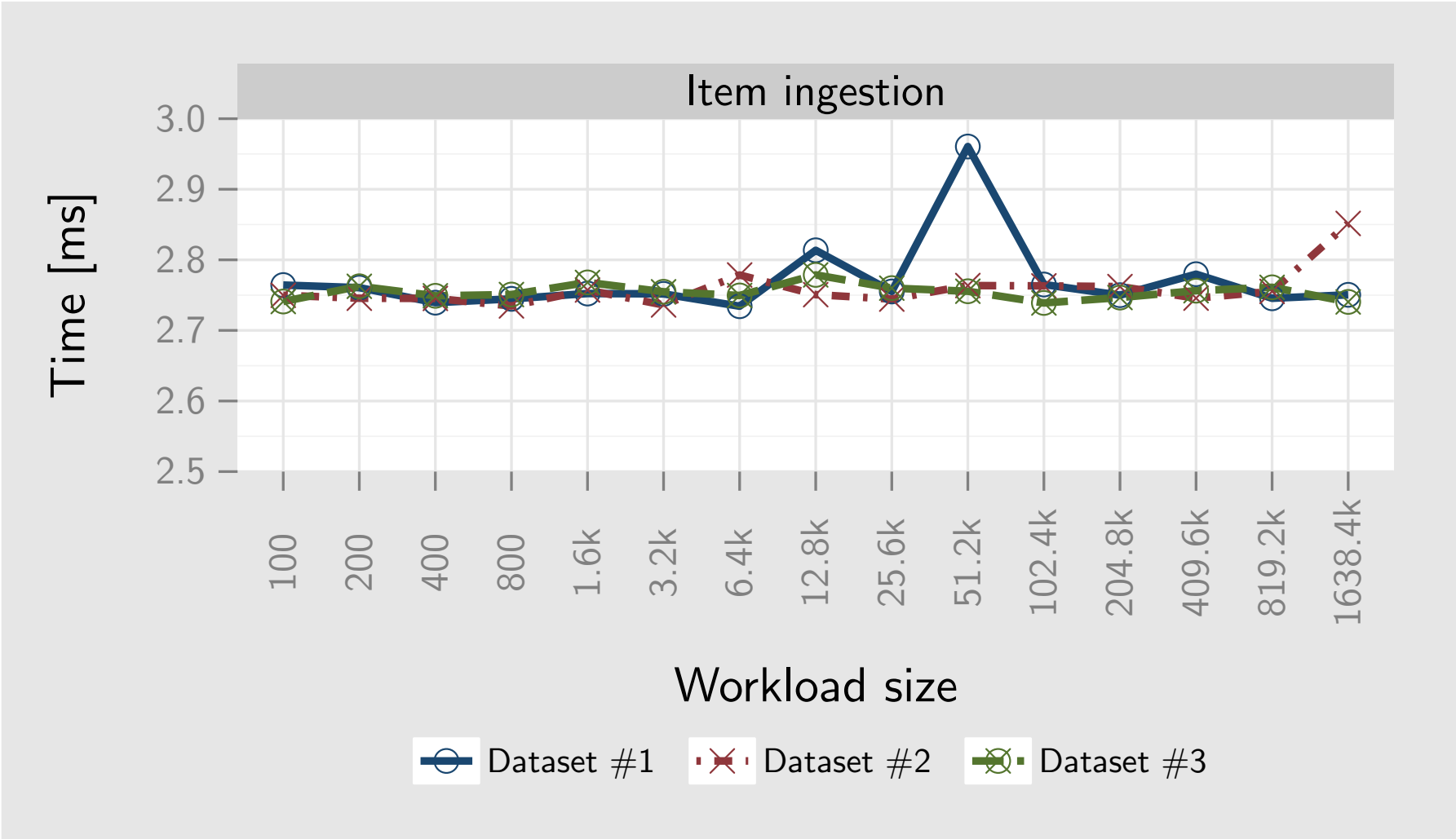
- Transaction log analysis —`http://pubs.cs.uct.ac.za`
  - Ingestion
  - Full-text search
  - Indexing operations
  - OAI-PMH data provider
  - Feed generation



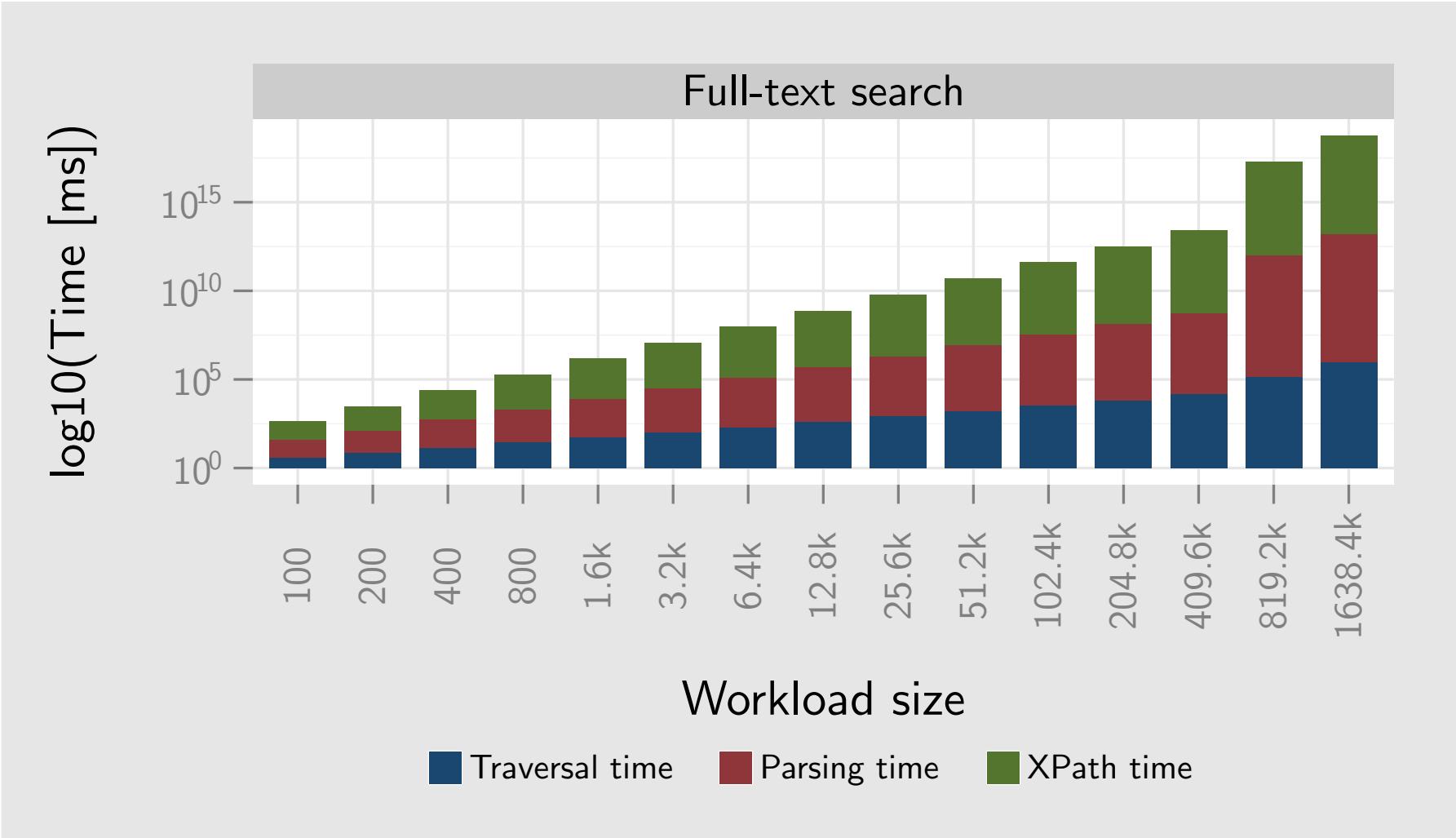
Performance experiment - experimental design

- Performance benchmarking
  - Evaluation aspects
  - Three-run averages for all scenarios
    - Datasets #1, #2 and #3
    - 15 workloads
  - Break-even points for performance degradation
  - Nielsen's three important limits for response times
- Performance comparisons
  - Benchmark results vs DSpace 3.1
    - Ingestion
    - Full-text search
    - OAI-PMH data provider

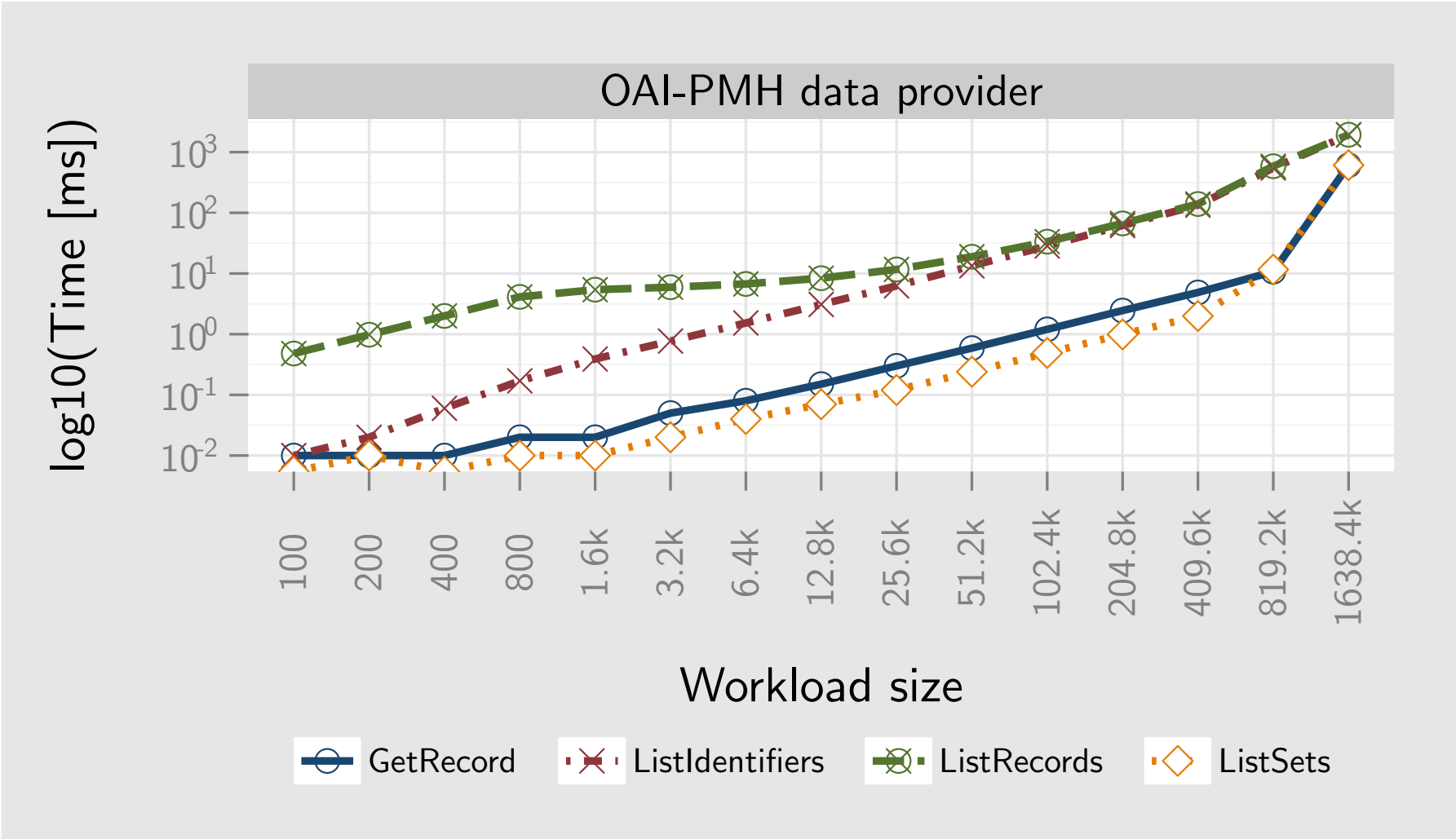
Performance experiment - results



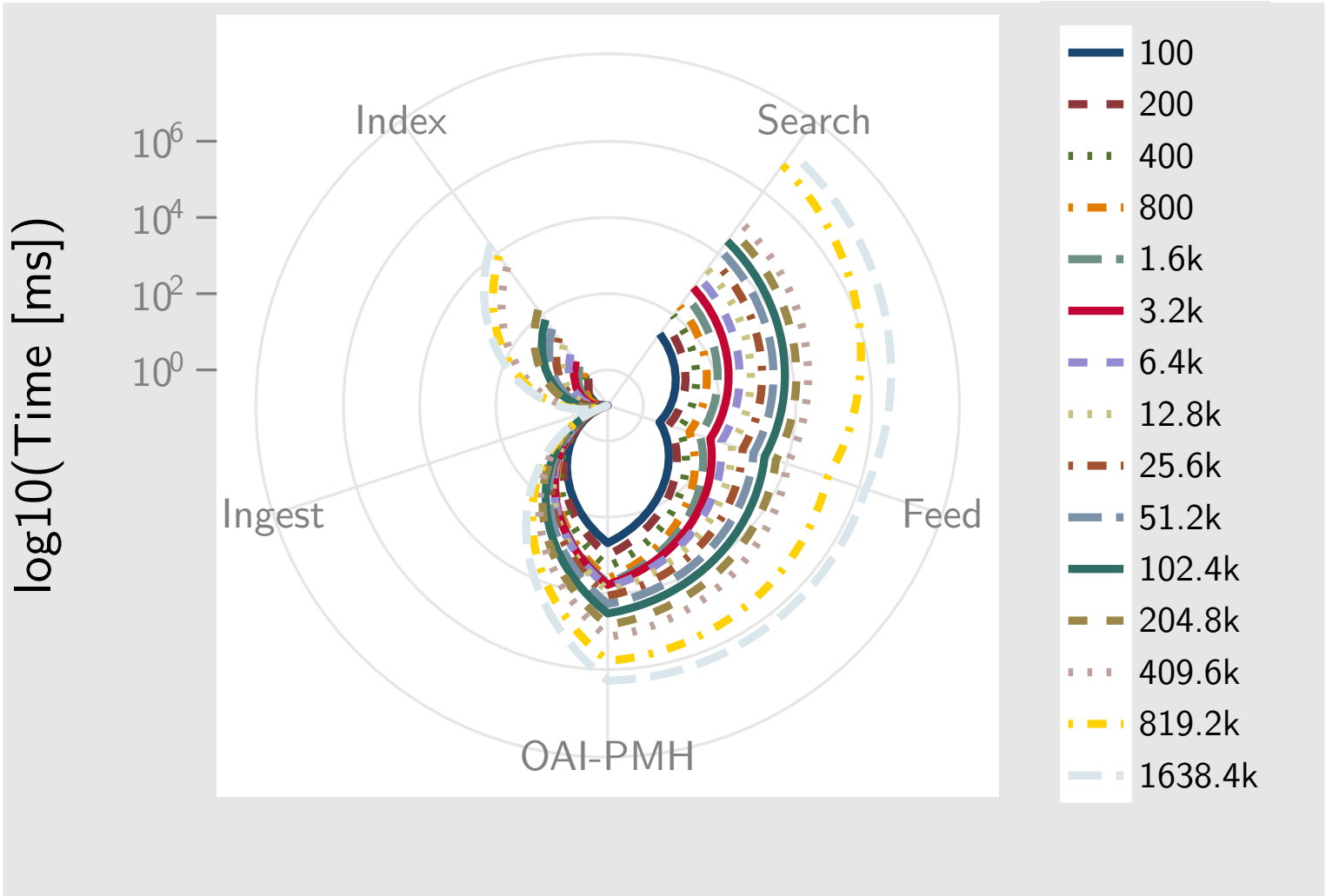
Performance experiment - results (2)



Performance experiment - results (3)



Performance experiment - results (4)



Performance experiment - findings

- Performance benchmarking
  - Performance within 'acceptable' limits for medium-sized collections
  - Ingestion performance NOT affected by collection scale
  - Performance generally degrades for collections $> 12\,800$ objects
  - Performance degradation adversely affects information-discovery services —Feed generation, full-text search and OAI-PMH data provider
- Comparison with DSpace 3.1
  - Ingestion performance better than DSpace
  - Information discovery operation —search and OAI-PMH— are slower than DSpace
    - DSpace uses Apache Solr for index
    - Comparable speeds can be attained through integration with third-party search services

Conclusions and future work

- Conclusions
  - Feasibility of simple DL architectures
  - Simplicity does not affect flexibility and potential extensibility of result tools and services
  - Performance acceptable for small- and medium-sized collections
  - Comparable features with well-established solutions
- Reference implementation
  - Packaging
  - Version control

Thank You

Questions?

Additional Information

http://dl.cs.uct.ac.za