Simplicity as an enabler for digital humanities research in low-resource environments

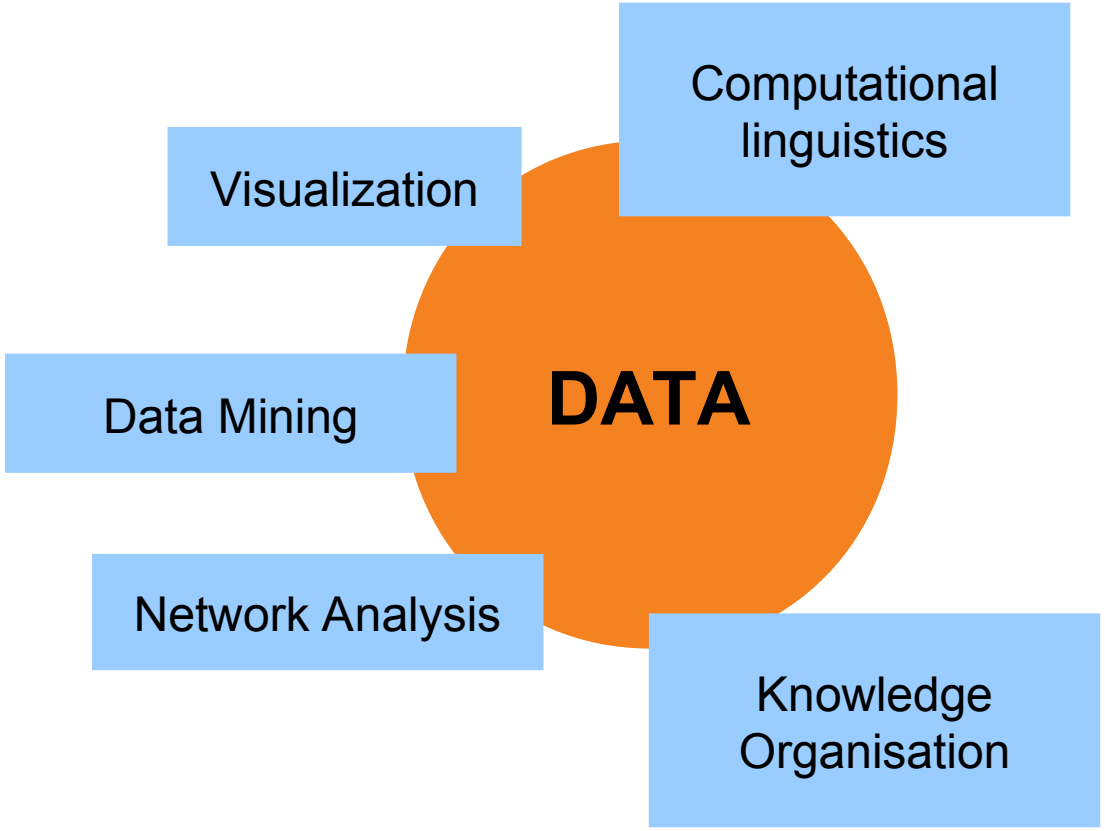
Hussein Suleman <hussein@cs.uct.ac.za>

Digital Libraries Laboratory
Department of Computer Science
School of IT
University of Cape Town
November 2021

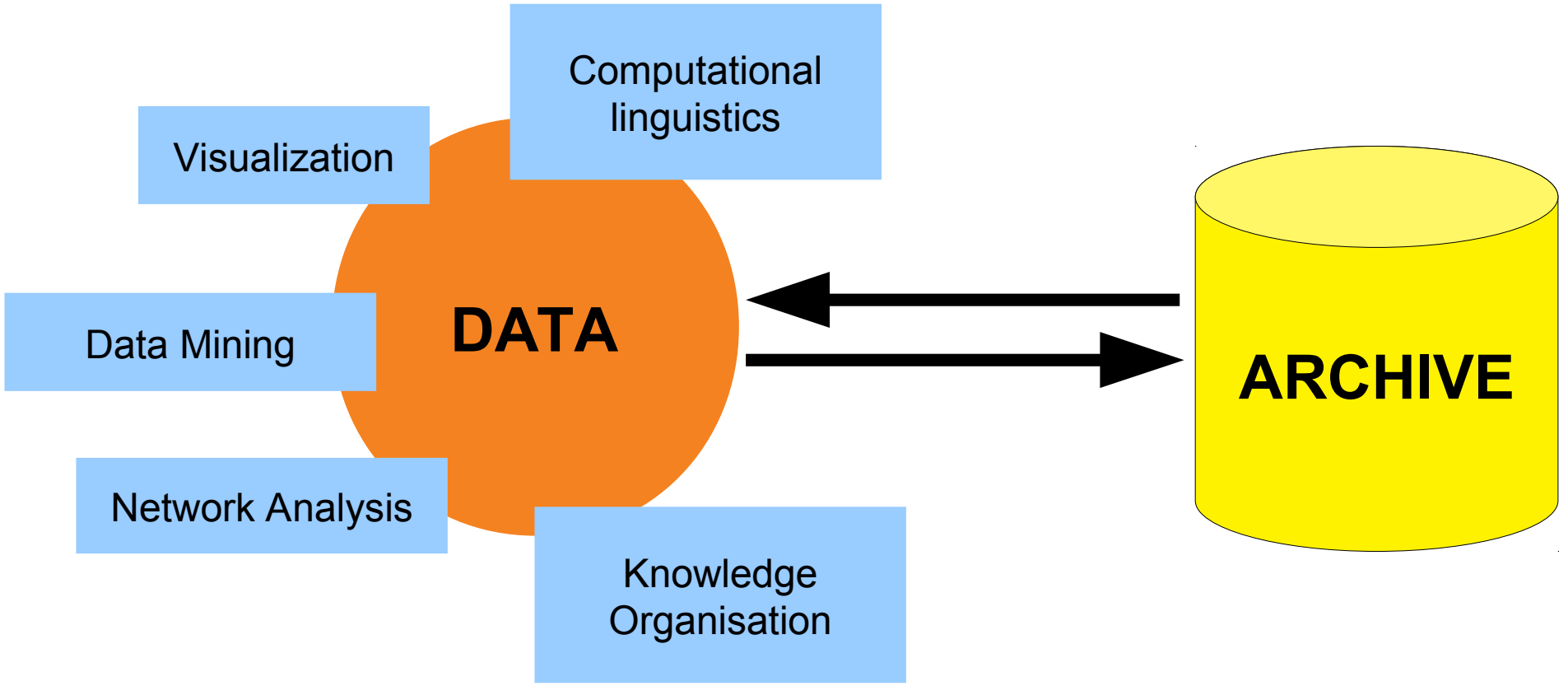
Last edited: 29 November 2021

Slide count: 39

Digital Humanities



Digital Humanities and Archives



Archives and Repositories

- Repositories are places to put stuff.
- Archives are long-term preservation mechanisms.

- Blurred distinction...

- Archives are a critical component of all research:
  - Primary datasets
  - Publications – Open Access
  - Theses and dissertations

South African Digitization Initiative (SADI)

- Going back about 10 years ...
  - Collaboration of government agencies (NRF, DST, Nat Lib, Nat Archives, ...), museums (Iziko, ...) and universities (UCT, UP, Wits, ...).
  - Build supporting infrastructure for research, especially in humanities.
  - Archives everywhere + central portal.

- Europeana sent people to help.
- Some interested external funders.
- But no progress ...

Resources to build archives

Money

Time

People

Institution

Low-Resource Environments

- Poor countries
  - Most of Africa.
- Specific areas
  - Rural areas all over the world.
- Specific organisations
  - NGO/NPOs relying on donor funding all over the world.
- Specific projects
  - Archivists who get funding for digitization but not maintenance.

Impact of low-resource environment

- Archives that are shut down (thesis archives in Africa).
- Archives at risk of being shut down.
- Archive divide (rich vs. poor).
- External providers (@mire).
- No archive.

What we need ...

- Archiving
  - Creation of metadata and digital objects.
  - Management and sharing of collections.
- Preservation
  - Maintaining digital collections in perpetuity.
  - Active curation of software/hardware/data.
- Data Access
  - To enable research in DH.

- *Without the drama of funding ending, staff shortages and access issues ... and collections disappearing or never starting...*

How to overcome low-resource issues?

- Clearly, custom solutions are not working.
- OSS is not always working for us either.
  - How are repository tools designed?
  - What are the principles governing this?
- Can we create a more suitable architecture for low resource environments?

Bleek and Lloyd Collection as exemplar?

THE DIGITAL BLEEK AND LLOYD HOME

This digital publication is part of a Llarec project to digitise, research and publish the Bleek and Lloyd Archive. The Digital Bleek and Lloyd includes scans of every page of the 110 Lucy Lloyd |xam notebooks, 17 Lloyd (mostly) !kun notebooks and 28 Wilhelm Bleek |xam notebooks. It also includes Jemima Bleek's solitary Korana and !kun notebook and four Lloyd Korana notebooks in the Maingard collection of the Library at the University of South Africa, as well as Dorothea Bleek's 32 notebooks. All the drawings and watercolours made by |han≠kass'o, Dia!kwain, Tamme, |uma, !nanni and Da are also in the digital collection. The digital archive includes a 280 000-word searchable index, cross-referenced and including notes and summaries for each of the stories listed. Notes in italics are direct quotes from the reports of Bleek and Lloyd in which they detailed the progress of their research.

Llarec (the Lucy Lloyd Archive, Resource and Exhibition Centre) is part of the Centre for Curating the Archive, a University of Cape Town research centre directed by Pippa Skotnes and located at the Michaelis School of Fine Art. The initial "Digital Bleek and Lloyd" accompanied the publication "Claim to the Country: the Archive of Wilhelm Bleek and Lucy Lloyd" by Pippa Skotnes (2007), published by Jacana Media and Ohio University Press. Subsequently Jemima Bleek's and Dorothea Bleek's notebooks have been added, as well as the Digital Stow, featuring the rock art copies of George Stow. The search index and summaries have also been extended and currently the Bleek and Lloyd dictionaries are being digitised. Please refer to the CCA website at http://www.cca.uct.ac.za for updates.

The project has been made possible by funding provided by the Andrew W. Mellon Foundation and De Beers; and is the result of the cooperation of the four curating institutions: University of Cape Town, Unisa, Iziko South African Museum and The National Library of South Africa.

These scans of the documents and images that comprise the Bleek and Lloyd archive may not be used or reproduced for any purpose without permission of the copyright holders.

Other Prior Work

- Greenstone – archives distributed on CDROM.
- Project Gutenberg – long term preservation using simple formats.
- LOCKSS – enable rescue by making many copies.

Principles and Design Goals 1/2

- **Simplicity/Minimalism** – only as much system as needed
- **No imposition on users** – no logins unless necessary
- **No Internet** - should work even when offline
- **Simple preservation** – as simple as copying files
- **Flexibility in objects/services** – no arbitrary restrictions
- **Superimposed and hierarchical information** – because we understand this
- **Platform agnosticism** – any OS, any hardware
- **Pre-processing** – fewer cycles needed when live

Principles and Design Goals 2/2

- It may be naieve to assume perfect digital preservation.
- Remember that preservation is
  - Informed by evolving research,
  - Actually unproven and unprovable,
  - And implemented by practitioners.

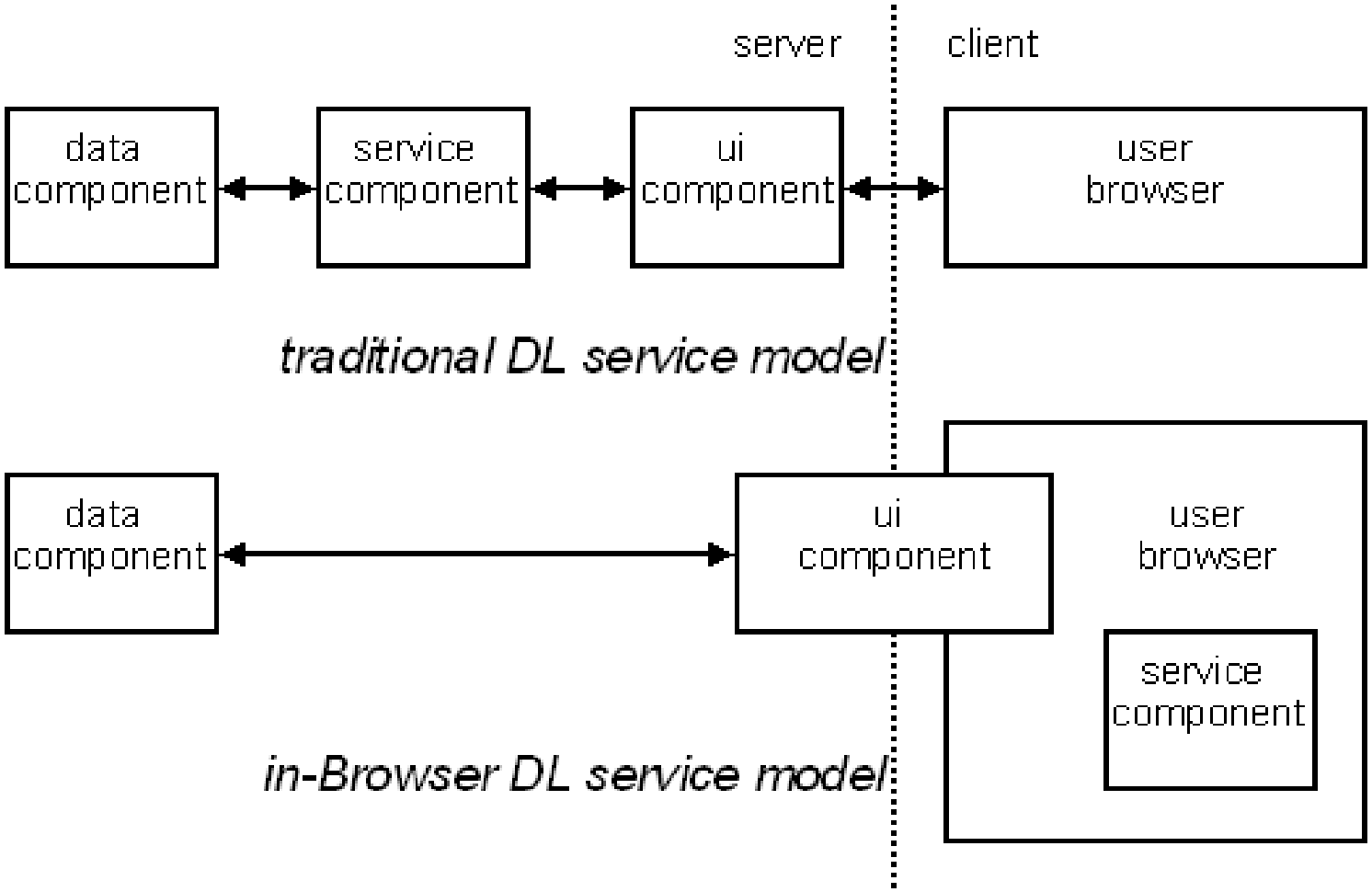
- **Design for Rescue**
  - What design decisions enable cheaper/faster/easier rescue after a disaster?
  - This was the philosophy behind LOCKSS.

Technical Options: Alternative Data Model

- XML/XSLT
  - Store all metadata/data in simple files in hierarchical directories. No database needed.

- Static generation
  - Site pre-generation (source->metadata->site) so software is not a long-term risk.

Technical Options: Alternative Service Model



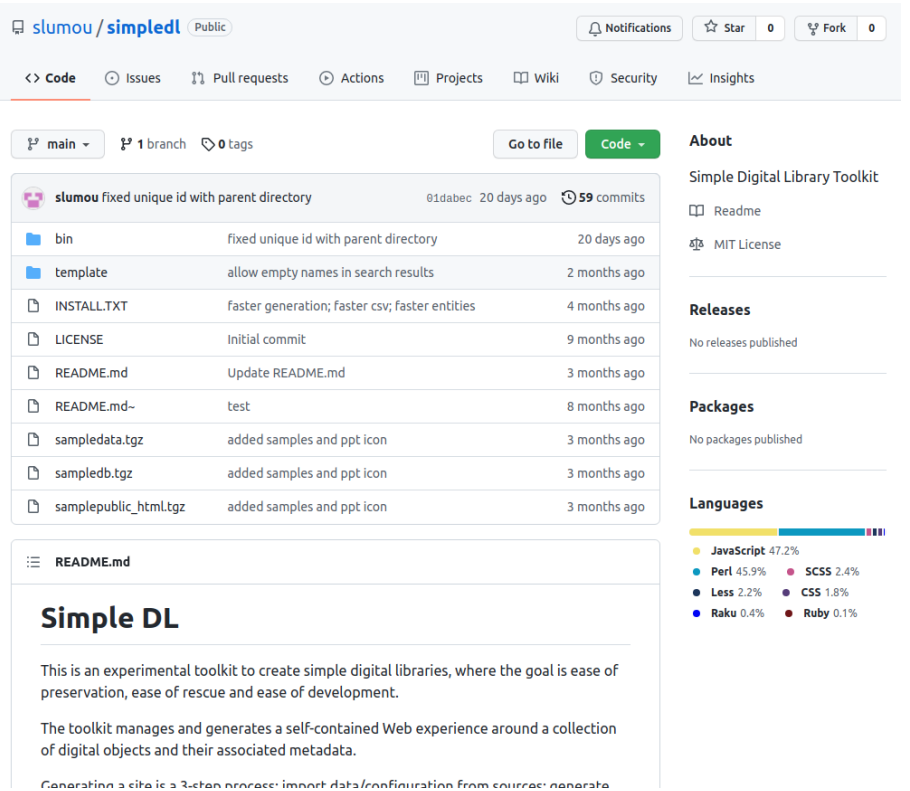
Simple Archives

- Proposition:
  - If we simplify the design of archives ...

- It will cost less to set up and maintain.
- It will be faster to create.
- It can be maintained without skilled people.
- It can be migrated easily.

Simple DL

- Toolkit to create simple archives.
- Based on many successful ideas in DL architecture, but built for low-resource environments and for adaptation.
- Experimental but:
  - Can change how we develop systems.
  - Provides a platform to experiment.

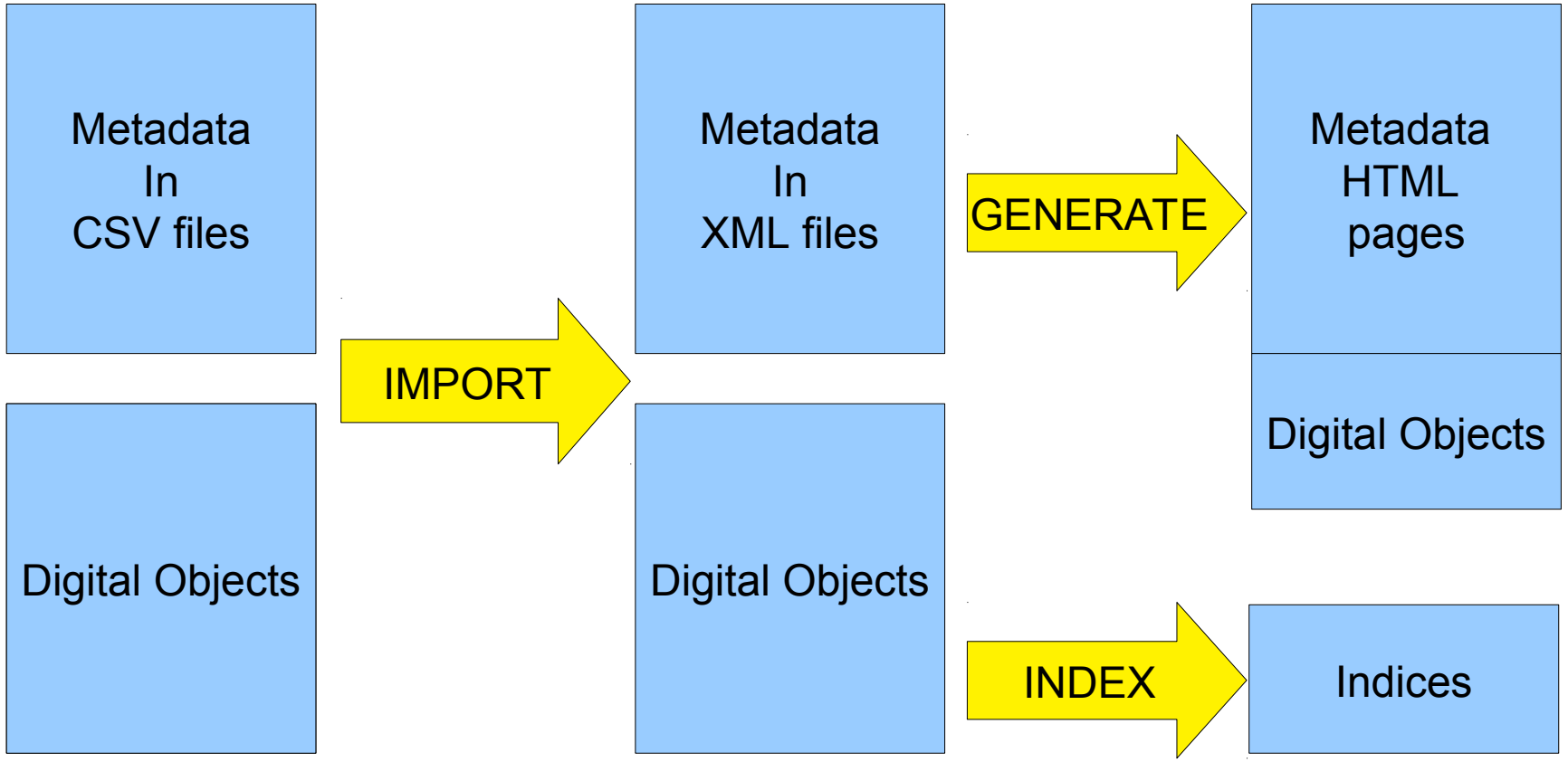


https://github.com/slumou/simpledl

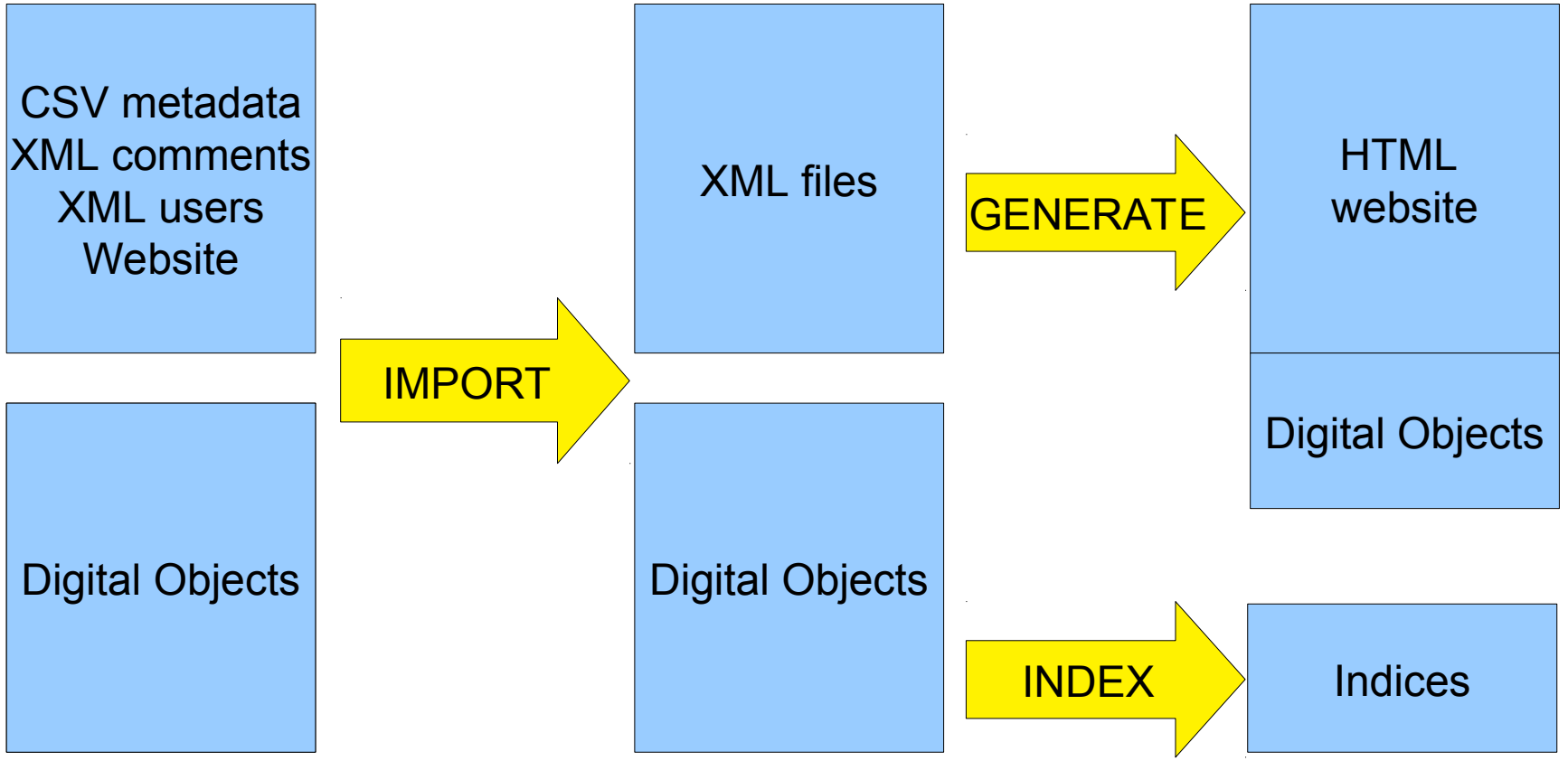
Major Features

- Offline pre-generated core website.
- Search/Browse in Javascript.
- Metadata stored in spreadsheets and XML files.
- Any metadata format (DC, ICA-Atom).
- Minimal use of Web applications.
- XSLT + CSS for styling.
- User and entity management.
- Submissions, comments, moderation.

Basic Idea


Metadata
In
CSV files
Digital Objects
IMPORT
Metadata
In
XML files
Digital Objects
GENERATE
Metadata
HTML
pages
Digital Objects
INDEX
Indices

Basic Idea - Extended

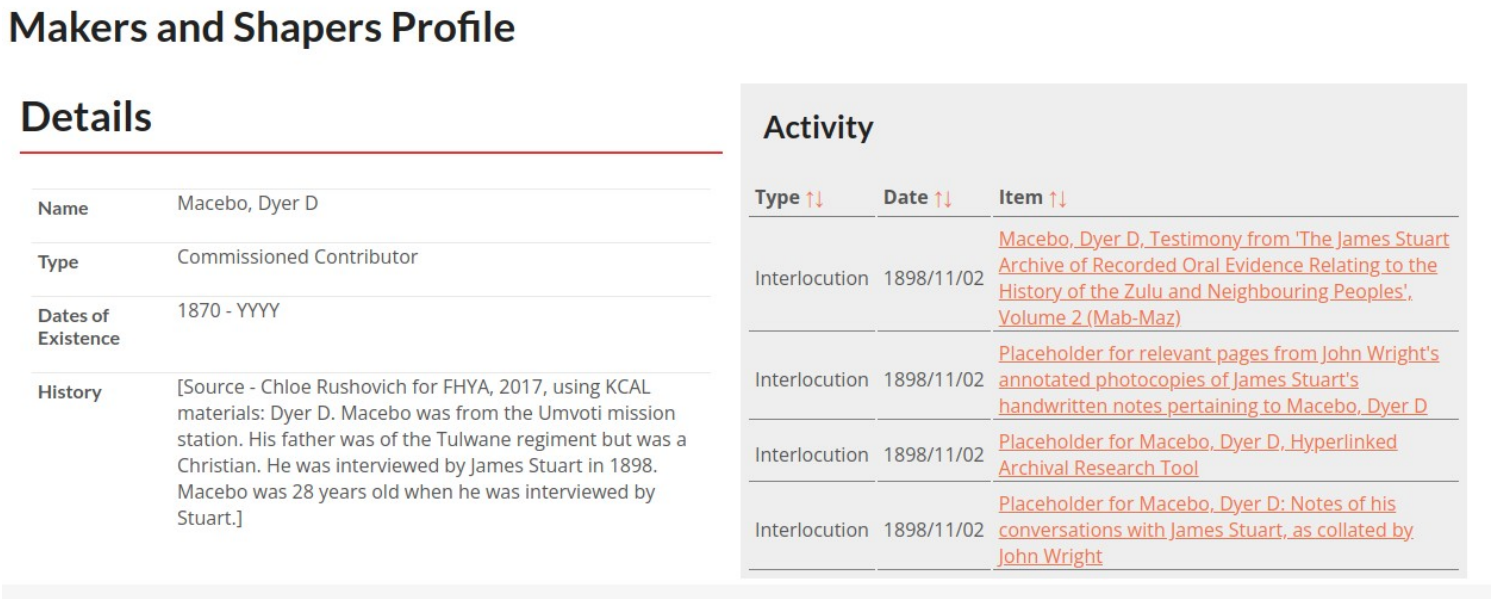


Management Interface

- Web UI to manage files, moderate submissions and import/generate/index.

Users and Entities

- Users have accounts on the system.
- Entities are extracted from metadata.
- Every entity page can links to authority records and items.



Makers and Shapers Profile

Details

Name	Macebo, Dyer D
Type	Commissioned Contributor
Dates of Existence	1870 - YYYY
History	[Source - Chloe Rushovich for FHYA, 2017, using KCAL materials: Dyer D. Macebo was from the Umvoti mission station. His father was of the Tulwane regiment but was a Christian. He was interviewed by James Stuart in 1898. Macebo was 28 years old when he was interviewed by Stuart.]

Activity

Type ↑↓	Date ↑↓	Item ↑↓
Interlocution	1898/11/02	Macebo, Dyer D, Testimony from 'The James Stuart Archive of Recorded Oral Evidence Relating to the History of the Zulu and Neighbouring Peoples', Volume 2 (Mab-Maz)
Interlocution	1898/11/02	Placeholder for relevant pages from John Wright's annotated photocopies of James Stuart's handwritten notes pertaining to Macebo, Dyer D
Interlocution	1898/11/02	Placeholder for Macebo, Dyer D, Hyperlinked Archival Research Tool
Interlocution	1898/11/02	Placeholder for Macebo, Dyer D: Notes of his conversations with James Stuart, as collated by John Wright

Submissions and Comments

- Users can add comments and attachments; or make new submissions.

Contributions

Name: ettore morelli

Date: 2021-06-24T14:30:44+0200

This research note is best read together with the note written by John Wright on the same subject. The two research notes were produced during the creation of the EMANDULO presentation 'Exploring uMgungundlovu' between the end of 2020 and the beginning of 2021.

The research note by Wright is here:
http://emandulo.apc.uct.ac.za/metadata/uMgungundlovu/28/1352/index.html

And the presentation 'Exploring uMgungundlovu' is here:
http://emandulo.apc.uct.ac.za/cgi-bin/view/Presentations/httpsstudio-emandulouctaczafhya-exploring-umgungundlovu.zip/studio-emandulo.uct.ac.za/https://studio-emandulo.uct.ac.za/fhya-exploring-umgungundlovu/index.html

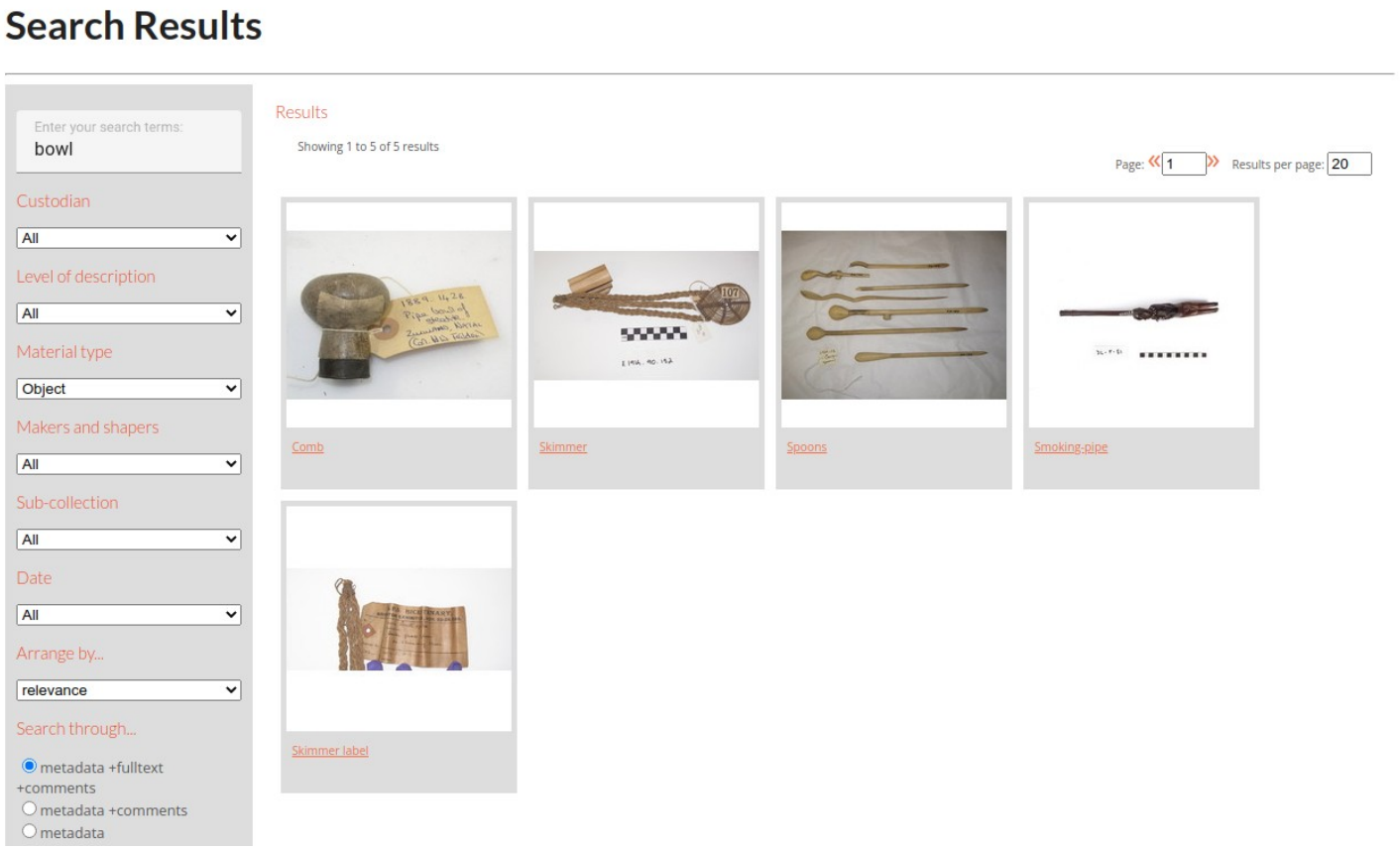
They both relate to the edited and interactive map of uMgungundlovu that is here attached in static form.

Attached File:

Exploring uMgungundlovu Map Static

Search and Browse

- Faceted search implemented in Javascript.

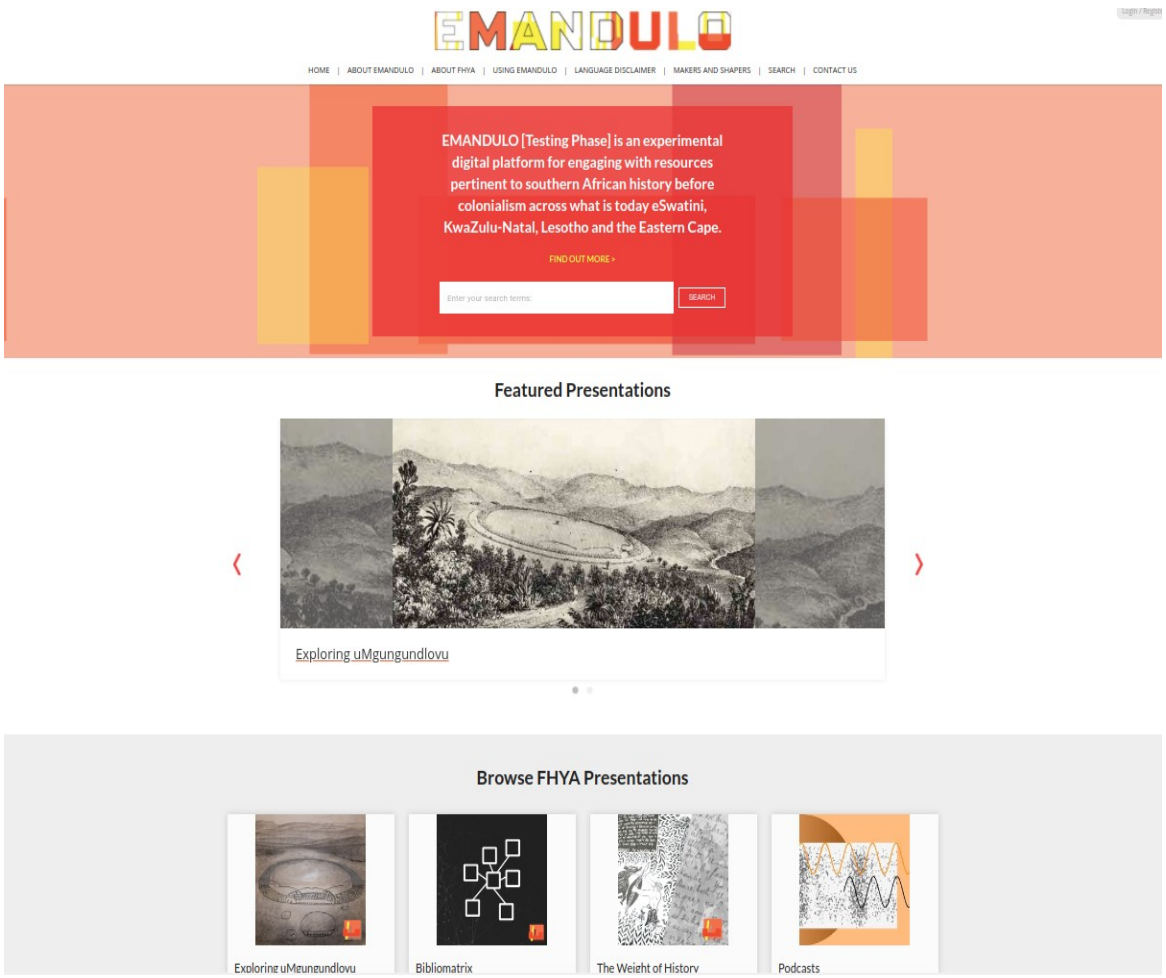


Evaluation

- Scalability:
  - Advanced faceted search queries:
    - Multiple search terms + multiple browse facets
    - 30000 items searched in 0.3s
  - Up to 100000 items handled in sub-second times

- Case Studies:
  - Emandulo, Re-source
  - New Bleek and Lloyd
  - NDLTD Document Archive, CS-HEAP

Case Study 1: Emandulo 1/2



Case Study 1: Emandulo 2/2

EMANDULO

HOME | ABOUT EMANDULO | ABOUT FHYA | USING EMANDULO | LANGUAGE DISCLAIMER | MAKERS AND SHAPERS | SEARCH | CONTACT US

Home
↳ Early KZN Materials collected by James Stuart and associated items
↳ FHYA selection of items pertinent to the testimonies of individual interlocutors from six published volumes of the James Stuart Archive
↳ FHYA selection of items pertinent to the testimonies of individual interlocutors from Volume 1 of the James Stuart Archive (Ant-Lyl)

FHYA selection of items pertinent to the testimonies of individual interlocutors from Volume 1 of the James Stuart Archive (Ant-Lyl)

Scope and Content

[Source - Debra Pryor for FHYA, 2019: Series contains front and back matter and the individual testimonies from "The James Stuart Archive of Recorded Oral Evidence Relating to the History of the Zulu and Neighbouring Peoples", Volume 1 (Ant-Lyl), edited by Colin de B. Webb and John Wright, together with their related material.

The volume was edited in the period April 1971 to November 1975. Wright was responsible for translating into English words and passages recorded by Stuart in isiZulu. He discussed points of translation with Professor A.T. Cope, head of the Department of Bantu Languages at the University of Natal in Durban. The volume was published in March 1976. It has since been reprinted.]

Metadata

Title	FHYA selection of items pertinent to the testimonies of individual interlocutors from Volume 1 of the James Stuart Archive (Ant-Lyl) [Source of title : FHYA]
Material Designation	Textual record
Reproduction Conditions	Creative Commons License: CC BY-NC-ND https://creativecommons.org/licenses/by-nc-nd/3.0/ Unless otherwise stated the copyright of all material on the FHYA resides with the contributing institution/custodian.
Events	

Contents

File: James Stuart Archive, Volume 1, Front matter

File: Antel, Mr., - Interview and associated items

File: Baleka ka Mpitikakazi - Interview and associated items

Case Study 2: New Bleek and Lloyd 1/2

Home
↳ Stories
↳ Wilhelm Bleek notebooks
↳ Words and sentences (given by Adam Kleinhardt)

Words and sentences (given by Adam Kleinhardt)

Metadata

Title	Words and sentences (given by Adam Kleinhardt)
Collection	Wilhelm Bleek notebooks
Contributor	Adam Kleinhardt
Summary	Words and sentences given by Adam Kleinhardt.
Comments	1) This story contains many notes on the verso pages concerning names of ǀxam, their genealogies, the names of the places they come from and the names of different groups of ǀxam (for example, see p.5v or p.12v for the genealogy of Adam Kleinhardt). 2) Notes were

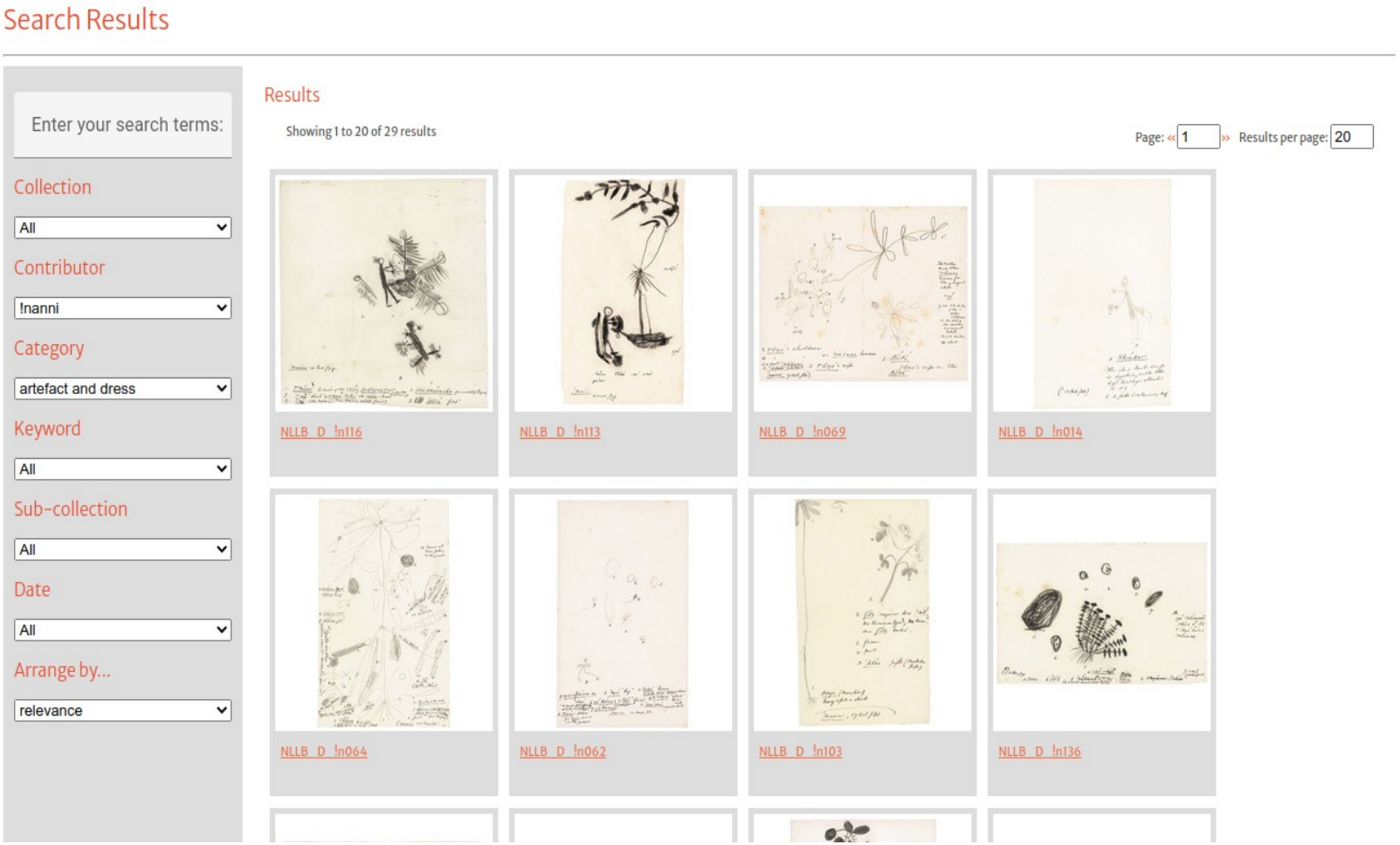
Contents

A1_4_1_00001.JPG

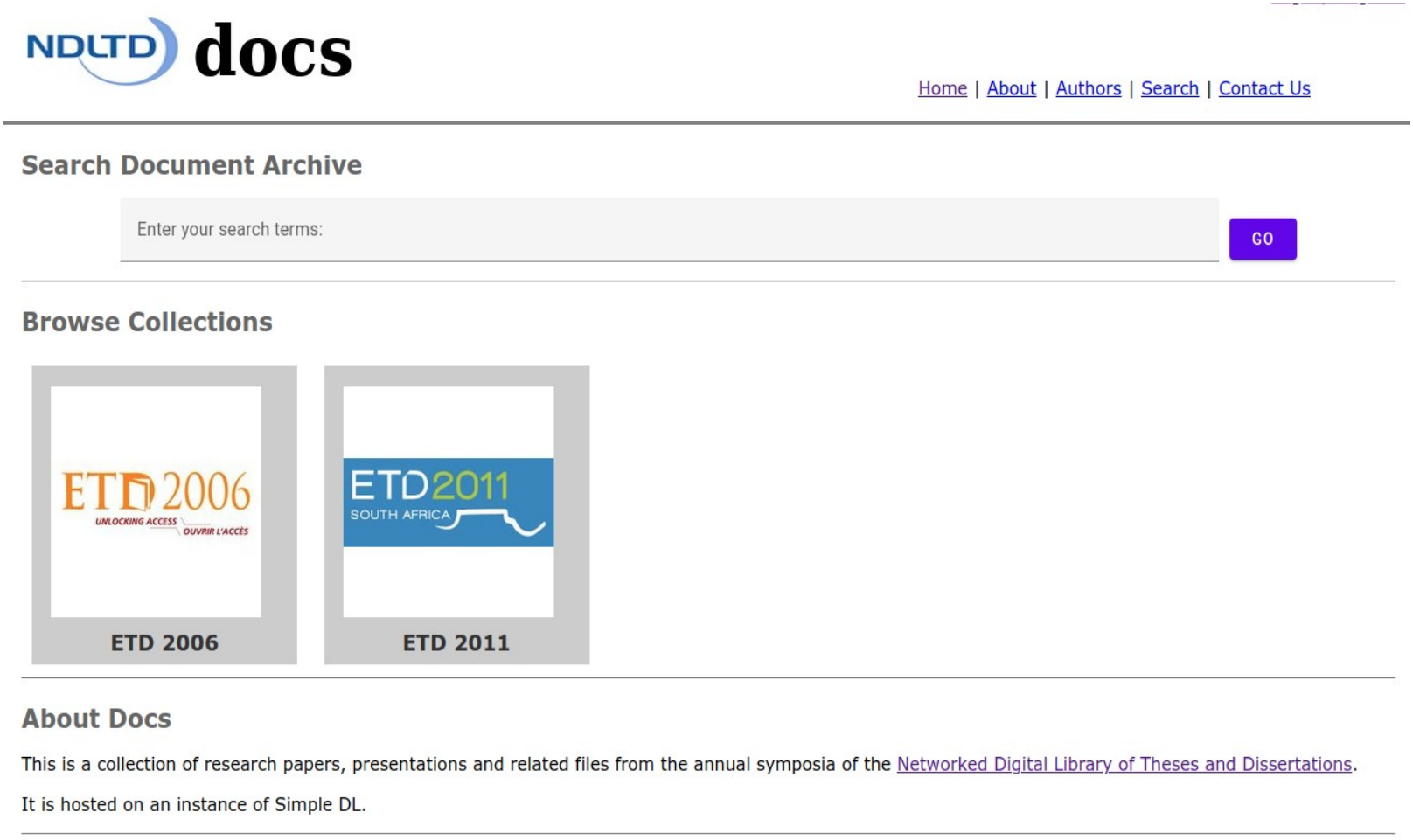
A1_4_1_00002.JPG

A1_4_1_00003.JPG

Case Study 2: New Bleek and Lloyd 2/2



Case Study 3: NDLTD Document Archive 1/2



Case Study 3: NDLTD Document Archive 2/2

NDLTD docs

Home | About | Authors | Search | Contact Us

Home
↳ ETD 2011
↳ Do ETDs deter publishers? Does web availability count as prior publication?

Do ETDs deter publishers? Does web availability count as prior publication?

Content

Metadata

Title	Do ETDs deter publishers? Does web availability count as prior publication?
Author	Marisa Ramirez Gail Mcmillan Joan Dalton Nancy Seamans Max Read

The Bootstrapping Philosophy

- This is a bootstrapping project.
  - Quickly create new archives for small collections in low-resource environments.
  - Meet the needs of most of the people most of the time.
  - Support experimentation and customisation.
  - Support eventual transition to other systems.

- This is not a solution for all problems.
  - One size does not fit all.
  - One design/architecture does not meet all needs.

Ongoing Work

- Resolvable unique identifiers – using purls.
- Complex object support – archiving entire presentations/exhibitions.
- Archive archiving – to manage versions of evolving knowledge structures.
- Virtualisation – to create/manage simple archives remotely.
- Complex object authoring tools – to create visualisations of concepts maps.
- Bookmarking facilities – both personal and community-oriented.

Beyond Archives

How does simplicity as a concept impact on the design of other tools used in DH?

Some Reflections 1/2

- How people engage with data is a rapidly evolving and unsettled idea.
- Context matters and bias is real.

- An archive is not the final location for an object.
- An archive is one temporary location for an object.

- We are not all online or online all the time.

Some Reflections 2/2

- This is about compromises:
  - Archiving/Preservation are simpler.
  - Scalability is not as good (if you have millions of objects).
- This might help us get to a national portal faster.
- Thinking about resource limits suggests radically different system approaches.
  - This will ultimately benefit everyone in the world!

References

- Suleman, H. (2021) Simple DL: A toolkit to create simple digital libraries, to appear in Proceedings of Asian Digital Libraries Conference (ICADL), December 2021.
- Suleman, H. (2019) Investigating the effectiveness of client-side search/browse without a network connection, Proceedings of 21st International Conference on Asia-Pacific Digital Libraries (ICADL), 4-7 November 2019, Kuala Lumpur, Malaysia, Springer. Available: http://pubs.cs.uct.ac.za/id/eprint/1333/
- Suleman, H. (2019) Reflections on Design Principles for a Digital Repository in a Low Resource Environment, Proceedings of HistoInformatics Workshop 2019, 13 September 2019, Oslo, Norway, CEUR. Available: http://pubs.cs.uct.ac.za/id/eprint/1331/
- Phiri, L. and H. Suleman (2013). Flexible Design for Simple Digital Library Tools and Services. In Machanick, Philip and Mosiuoa Tsietsi, Eds. Proceedings SAICSIT '13 2013 South African Institute for Computer Scientists and Information Technologists, pages 160-169, East London, South Africa. Available http://pubs.cs.uct.ac.za/archive/00000891/
- Phiri, L., K. Williams, M. Robinson, S. Hammar and H. Suleman (2012) Bonolo: A General Digital Library System for File-based Collections. In Chen, Hsin-Hsi and Gobinda Chowdhury, Eds. Proceedings 14th International Conference on Asia-Pacific Digital Libraries 7634, pages 49-58, Taipei, Taiwan. Available http://pubs.cs.uct.ac.za/archive/00000784/
- Suleman, H., M. Bowes, M. Hirst and S. Subrun (2010), Hybrid Online-Offline Collections, in Proceedings of Annual Conference of the South African Institute for Computer Scientists and Information Technologists (SAICSIT 2010), Bela Bela, South Africa, 11-13 October 2010, ACM. Available http://pubs.cs.uct.ac.za/archive/00000655/
- Suleman, H. (2007). Digital Libraries Without Databases: The Bleek and Lloyd Collection", in Kovacs, Laszlo, Norbert Fuhr and Carlo Meghini (eds): Proceedings of Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007), pp. 392-403, 16-19 September, Budapest, Hungary, pp. 392-403, Springer. Available http://pubs.cs.uct.ac.za/archive/00000433/
- Suleman, H. (2007). in-Browser Digital Library Services, in Kovacs, Laszlo, Norbert Fuhr and Carlo Meghini (eds): Proceedings of Research and Advanced Technology for Digital Libraries, 11th European Conference (ECDL 2007), 16-19 September, Budapest, Hungary, pp.462-465, Springer. Available http://pubs.cs.uct.ac.za/archive/00000434/

Thank you for listening!

https://github.com/slumou/simpledl

Questions, comments and suggestions are welcome...

Contact me at: hussein@cs.uct.ac.za

Digital Libraries Laboratory
Department of Computer Science
School of IT
University of Cape Town