Information Retrieval and Data Mining for Development

Hussein Suleman

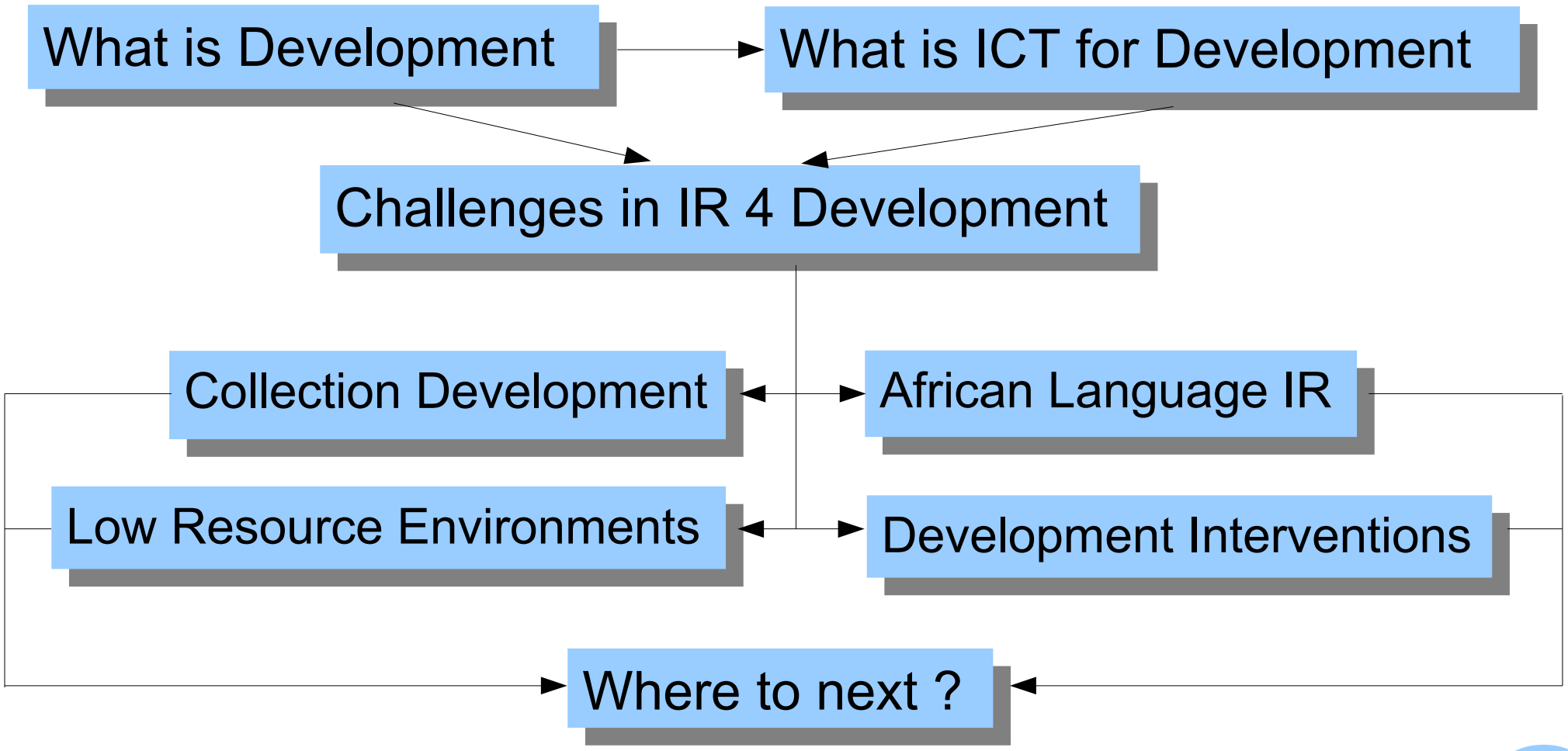
Digital Libraries Laboratory @ Centre for ICT4D
Department of Computer Science
School of IT
University of Cape Town

January 2020

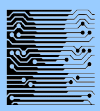
Key Research Question

How do we use
Information Retrieval / Data Mining /...
to support **Development** in Africa?

Outline of Talk



What is (Human/Socio-economic) Development?



Development Agendas

- **UN Millenium Development Goals**
- UN Millenium Declaration
- **UN Sustainable Development Goals**
- South Africa
  - **National Development Plan (2012)**
  - Growth Employment and Redistribution (1996)
  - Reconstruction and Development Plan (1994)
- Africa-wide
  - New Partnership for Africa's Development (NEPAD)
  - ...

UN Millenium Develpmemt Goals

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

SA National Development Plan 2012-2030

- The creation of jobs and the development of the economy
- Development of the economic infrastructure: coal and gas, water, electricity and telecommunications
- Environmental sustainability and management of environmental resources
- Development of an inclusive rural economy
- Regional and international trade
- Housing and urban/rural planning
- Education and training
- Medical care
- Safety and security
- Building capacity for a developmental state
- Fighting corruption
- Nation building for a unified society

Programme of the Austrian Federal Govt 2008-2013

Programme of
the Austrian Federal Government
for the 24th Legislative Period

Contents

Preamble Programme of the Austrian Federal Government 2008 – 2013
TOGETHER FOR AUSTRIA 3

EMPLOYMENT AND BUSINESS LOCATION POLICY 8

RESEARCH, TECHNOLOGY AND INNOVATION (RTI) 42

INFRASTRUCTURE AND TRANSPORT 51

AGRICULTURE AND FORESTRY, RURAL AREAS 69

CLIMATE AND THE ENVIRONMENT 80

HOME AFFAIRS, JUSTICE AND NATIONAL DEFENCE 95

SOCIETY, WOMEN, FAMILY AND EQUAL OPPORTUNITIES POLICY 154

SPORT 170

HEALTH AND SOCIAL POLICY 175

EDUCATION, SCIENCE, RESEARCH, CULTURE AND THE MEDIA 202

ART AND CULTURE 224

FOREIGN AND EUROPEAN POLICY 232

Nigeria Vision 20:2020

Table of contents

Introduction 4
Executive Summary 8

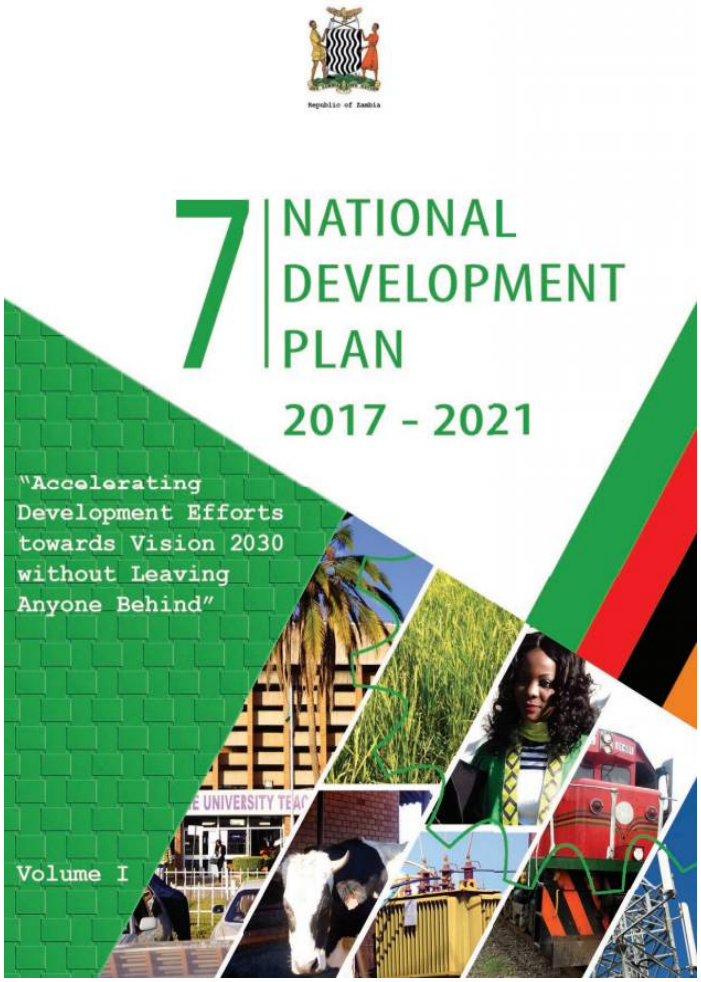
Section 1
The Vision and Development Priorities 14
Defining NV20:2020 14
The Vision and National Aspirations 14
The Strategic Framework for NV20:2020 15
Imperatives for Nigeria's Economic Transformation 16
The Burning Platform for Change 16
The Economic Growth Challenge 17
Domestic Constraints to Growth and Development 18
The Challenge of Growth in the Emerging Global Landscape 19
The Dream Economy-Our Economic Aspirations 21
Envisioned Economic Size & Structure 21
Macroeconomic Strategies And Policy Thrusts 22
Key Growth Drivers & Sectors Of Strategic Focus 24
Critical Policy Priorities 25

Section 2
Guaranteeing The Well-Being And Productivity Of The People 28
Eradicate extreme hunger and poverty 29
Enhance access to quality And affordable healthcare 31
Provide sustainable access to potable Water and basic sanitation 32
Provide accessible and affordable Housing 33
Build human capacity for sustainable Livelihoods and national development 34
Promote gender equality And empower women 38
Improve access to micro-credit 40
Foster a culture of entertainment and Recreation for enhanced productivity 41

Section 3
Optimising The Key Sources Of Economic Growth 48
Stimulate primary production To enhance the competitiveness Of Nigeria's real sector 48
Significantly increase production Of processed and manufactured Goods for export 52
Stimulate domestic and foreign Trade in value-adding products and services 59
Strengthen linkages between Key sectors of the economy 63

Section 4
Fostering Sustainable Social And Economic Development 70
Develop efficient, accountable, transparent and participatory governance 71
Establish a competitive, private Sector led business environment Characterized by sustained Macroeconomic stability 74
Enhance national security and Improve the administration of justice 76
promote unity in diversity, national pride, and the conservation of the nation's cultural heritage 77
Develop sufficient and efficient infrastructure to support sustained economic growth 78
Preserve the environment for sustainable socio-economic development 83
Promote the sustainable development of Nigeria's geo-political regions into economic growth poles 83

Zambia 7th National Dev Plan



10.3.2 Strategy 2: Expand capacity to increase access to quality health care ... 97
10.3.3 Strategy 3: Enhance food security and nutrition ... 98
10.3.4 Strategy 4: Promote private sector participation in health care delivery ... 98
10.3.5 Strategy 5: Accelerate human resource outputs, recruitment and retention ... 98
10.4 DEVELOPMENT OUTCOME 2: IMPROVED EDUCATION AND SKILLS DEVELOPMENT ... 99
10.4.1 Strategy 1: Enhance access to quality, equitable and inclusive education ... 100
10.4.2 Strategy 2: Enhance access to skills training ... 100
10.4.3 Strategy 3: Enhance private sector participation ... 101
10.4.4 Strategy 4: Continuous review of curriculum ... 101
10.4.5 Strategy 5: Enhance role of science, technology and innovation ... 101
10.5 DEVELOPMENT OUTCOME 3: IMPROVED ACCESS TO WATER SUPPLY AND SANITATION ... 102
10.5.1 Strategy 1: Enhance provision of adequate safe water and sanitation ... 102
10.5.2 Strategy 2: Improve availability of water and sanitation infrastructure ... 103
10.5.3 Strategy 3: Enhance research in water supply and sanitation services ... 103
10.5.4 Strategy 4: Promote alternative financing for water and sanitation ... 103
10.5.5 Strategy 5: Enhance provision of adequate solid waste management services ... 103
11. CREATING A CONDUCIVE GOVERNANCE ENVIRONMENT FOR A DIVERSIFIED AND INCLUSIVE ECONOMY ... 105
11.1 OVERVIEW ... 105
11.2 BINDING CONSTRAINTS TO A CONDUCIVE GOVERNANCE ENVIRONMENT FOR A DIVERSIFIED AND INCLUSIVE ECONOMY ... 105
11.3 DEVELOPMENT OUTCOME 1: IMPROVED POLICY ENVIRONMENT ... 106
11.3.1 Strategy 1: Reform and re-establish inclusive consultative structures ... 106
11.3.2 Strategy 2: Reform labour policy and legislative framework ... 106
11.3.3 Strategy 3: Accelerate implementation of the Decentralisation Policy ... 106
11.3.4 Strategy 4: Enhance national data and information systems ... 107
11.4 DEVELOPMENT OUTCOME 2: IMPROVED TRANSPARENCY AND ACCOUNTABILITY ... 107
11.4.1 Strategy 1: Strengthen transparency and accountability mechanisms ... 107
11.4.2 Strategy 2: Strengthen public financial management ... 108
11.5 DEVELOPMENT OUTCOME 3: AN INCLUSIVE DEMOCRATIC SYSTEM OF GOVERNANCE 108
11.5.1 Strategy 1: Promote platforms for citizen engagement and participation ... 108
11.5.2 Strategy 2: Strengthen electoral processes and systems ... 109
11.6 DEVELOPMENT OUTCOME 4: IMPROVED SERVICE DELIVERY ... 109
11.6.1 Strategy 1: Scale-up public service reforms ... 109
11.6.2 Strategy 2: Strengthen land administration and management ... 110
11.7 DEVELOPMENT OUTCOME 5: IMPROVED RULE OF LAW, HUMAN RIGHTS AND CONSTITUTIONALISM ... 110
11.7.1 Strategy 1: Strengthen prosecution system ... 110
11.7.2 Strategy 2: Enhance access to justice ... 111
11.7.3 Strategy 3: Promote human rights ... 111
11.8 DEVELOPMENT OUTCOME 6: ENHANCED NATIONAL VALUES, PRINCIPLES AND ETHICS ... 111
11.8.1 Strategy 1: Promote inculcation of national values and principles ... 112
11.8.2 Strategy 2: Promote platforms for religious engagement and participation ... 112
11.8.3 Strategy 3: Enhance research in application of values and principles ... 112

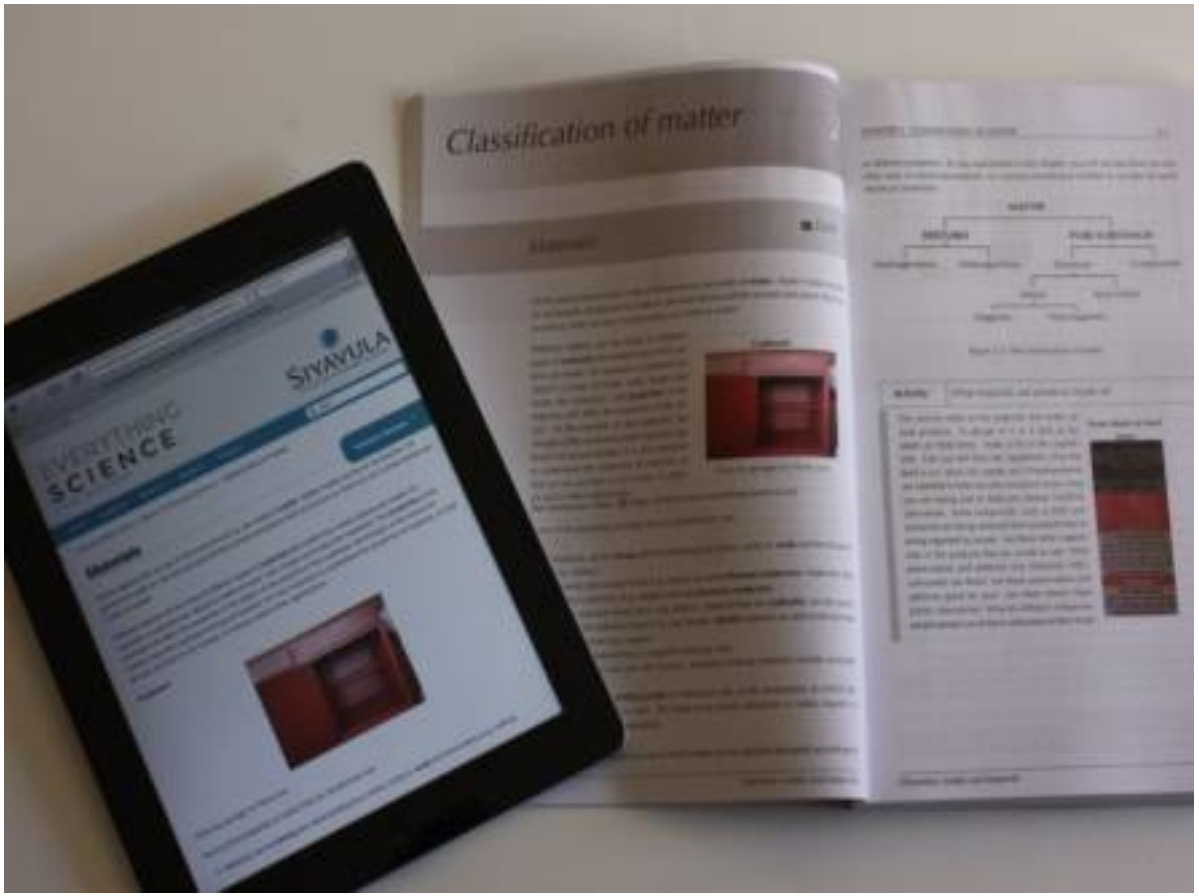
PART V: 7NDP FINANCING ... 115
12. FINANCING THE 7NDP ... 117
12.1 OVERVIEW ... 117
12.2 RESOURCE ENVELOPE ... 117
12.3 RESOURCE MOBILISATION STRATEGY ... 117
12.4 EXPENDITURE ... 118
12.5 EXTERNAL GRANTS (ASSUMPTIONS) ... 119
12.6 PRIVATE SECTOR FINANCING ... 119
12.7 DOMESTIC AND EXTERNAL FINANCING ... 119
12.8 OTHER FINANCING OPTIONS ... 119

The Decolonisation Debates

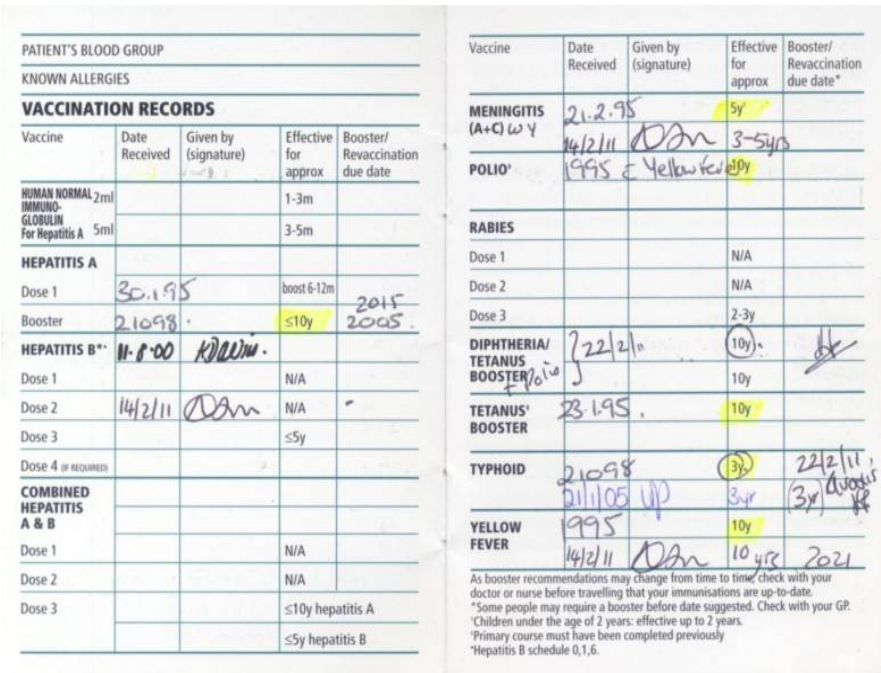
- How do we decolonise African society?
  - Different knowledge systems? ICT? Do we do ICT differently?
  - Do we need a programming language with keywords in isiZulu?
  - Do we teach programming in isiZulu?
  - Public intellectuals or universal scholars?
  - Excellence vs. Local Relevance
- Why is AFIRM mostly run by people from the Northern Hemisphere?
- What do they say: Ngũgĩ wa Thiong'o, Mahmood Mamdani,...

What is ICT for Development

What is ICT4D: Example 1/4



What is ICT4D: Example 2/4



PATIENT'S BLOOD GROUP

KNOWN ALLERGIES

VACCINATION RECORDS

Vaccine	Date Received	Given by (signature)	Effective for approx	Booster/ Revaccination due date
HUMAN NORMAL IMMUNO-GLOBULIN For Hepatitis A 2ml			1-3m	
5ml			3-5m	
HEPATITIS A				
Dose 1	30.1.95		boost 6-12m	
Booster	21098		≤10y	2015 2005
HEPATITIS B*	11-8-00	KDavin		
Dose 1			N/A	
Dose 2	14/2/11	[illegible]	N/A	
Dose 3			≤5y	
Dose 4 (IF REQUIRED)				
COMBINED HEPATITIS A & B				
Dose 1			N/A	
Dose 2			N/A	
Dose 3			≤10y hepatitis A	
			≤5y hepatitis B	

Vaccine	Date Received	Given by (signature)	Effective for approx	Booster/ Revaccination due date*
MENINGITIS (A+C) WY	21.2.95		5y'	
	14/2/11	[illegible]	3-5yrs	
POLIO'	1995 c Yellow fever		10y	
RABIES				
Dose 1			N/A	
Dose 2			N/A	
Dose 3			2-3y	
DIPHTHERIA/ TETANUS BOOSTER + Polio	22/2/11		10y	[illegible]
			10y	
TETANUS' BOOSTER	23.1.95		10y	
TYPHOID	21098		3yr	22/2/11
	21/1/05	UP	3yr	(3yr) Avoids
YELLOW FEVER	1995		10y	
	14/2/11	[illegible]	10 yrs	2021

As booster recommendations may change from time to time, check with your doctor or nurse before travelling that your immunisations are up-to-date.
*Some people may require a booster before date suggested. Check with your GP.
'Children under the age of 2 years: effective up to 2 years.
'Primary course must have been completed previously
*Hepatitis B schedule 0,1,6.

What is ICT4D: Example 3/4

What is ICT4D: Example 4/4

The Big Question

- Can we use ICT to aid human development?
- Can we use IR/DM to aid human development?

Challenges: IR/DM for Development

Goal: IR for Human Development

- Human Dignity
  - Promote the status of local languages.
  - Create tools that support local languages.
  - Increase presence of local languages.

- IR4D
  - IR for employment, governance, health, etc.

Challenge 1: IR algorithms

- Little algorithmic support in IR/NLP.

- Are there language-specific tools/algorithms in African languages?
  - How well do they work?
  - How many languages are supported?

Challenge 2: Data

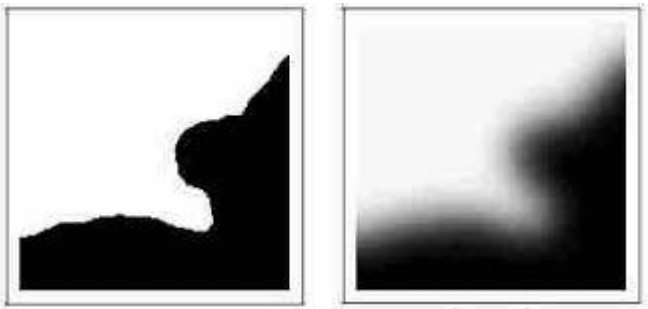
- Very little, mixed language and noisy data.
- <1000 Wikipedia documents for some African languages.
- Few corpora!
  - How much electronic content do we produce?

№	Language	Language (local)	Wiki	Articles
254	Tulu	ತುಳು	tcy	937
255	Cherokee	ᏣᎳᎩ	chr	854
256	Latgalian	Latgaļu	ltg	806
257	Samoan	Gagana Samoa	sm	797
258	Oromo	Oromoo	om	772
259	Ingush	Гӏалгӏай	inh	764
260	Xhosa	isiXhosa	xh	737
261	Old Church Slavonic	Словѣньскъ	cu	657
262	Romani	romani - रोमानी	rmy	657
263	Bambara	Bamanankan	bm	646
264	Tswana	Setswana	tn	641
265	Norfolk	Norfuk	pih	639
266	Kirundi	Kirundi	rn	611
267	Cheyenne	Tsetsêhestâhese	chy	609
268	Twi	Twi	tw	606
269	Gothic	□□□□□□	got	563
270	Tumbuka	chiTumbuka	tum	562
271	Tsonga	Xitsonga	ts	561
272	Akan	Akana	ak	559
273	Sesotho	Sesotho	st	539
274	Atikamekw	Atikamekw	ati	500

Challenge 3: Fuzziness

- Unclear language boundaries.

- How many languages are there?
  - How many have been clearly defined?
  - How many are managed?
- What is a language and what is a dialect/accent?

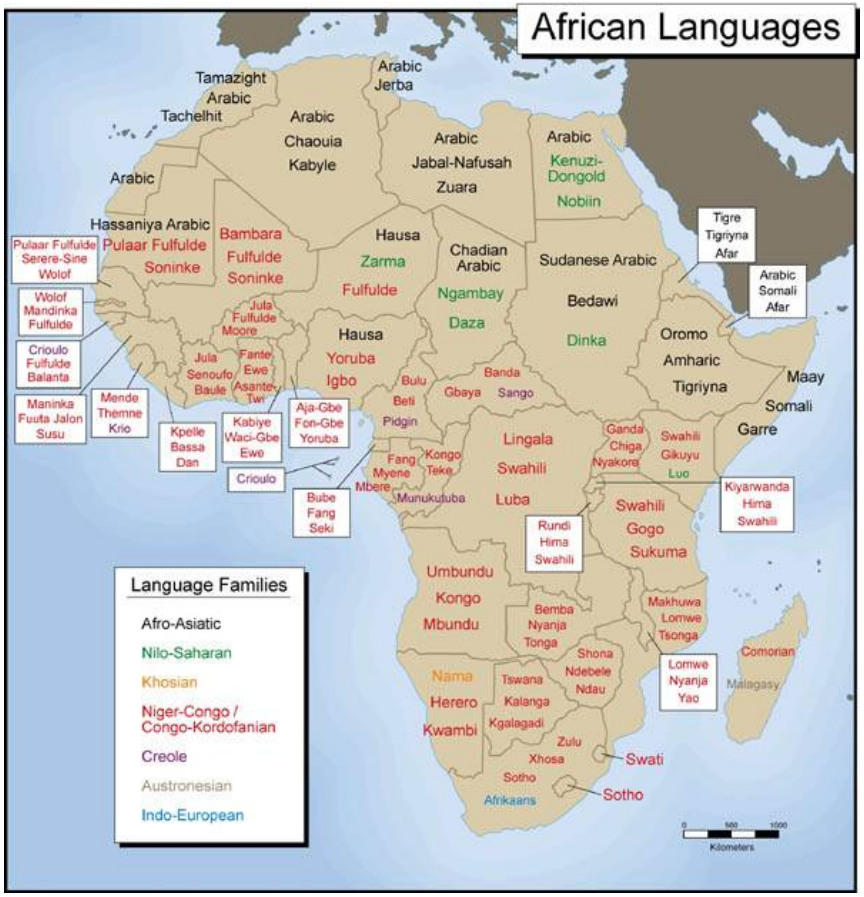


Challenge 4: Digital Divide

- Access / Knowledge

- How many people understand how to search?
- How many people use search?
  - Do people even have Internet access?

Challenge 5: Many Languages



- Multilingualism is the norm.
- How many languages do people use?
- Are documents/queries in one language or are they mixed?

Challenge 6: Resource Limits

- We do not have the resources.

- Limited skills among researchers.
- Limited bandwidth to access data.
- Limited skills among users.
- Limited funding for anything.

Studies: Collection Development

Corpora: Language Detection

Meluleki Dube, U/G

- Can we successfully determine the language, from among a group of 9 related African languages, of a piece of text?
  - Web page?
  - Tweet?
- Trigram modelling and model alignment distance gives up to 92% accuracy.
  - Incorrect predictions scatter by language similarity.

Corpora: Crowdsourcing

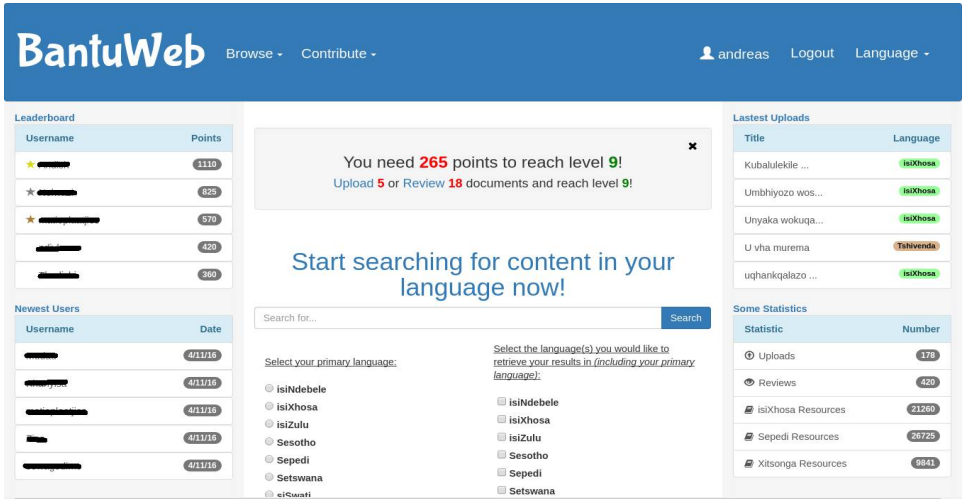
Sean Packham, MSc

- Parallel corpus in isiXhosa-English.
- Will people contribute if money paid is varied or there is no money but only gamification?
  - Payment is only criterion!

Corpora: SALANG

Andreas von Holy, Osher Shuman, Alon Bresler, Bsc(Hons)

- Create a central portal for documents in any SA Bantu language, with gamification, multilingual search, etc.



Corpora: Long-term effects

Jackson Moji, MSc (current)

- Does gamification for corpus creation work in the long term?
  - Will people lose interest?
  - Will they continue to contribute?
  - How is intrinsic motivation affected by time?

- Extension of SALang project.

Studies: African Language IR

Mixed Language IR

Mohammed Mustafa Ali, PhD

- Noted that Google is language unaware.
- Poor results for mixed queries – queries in multiple languages.
  - Dominant languages are dominant in results.
  - Mixed language use is very popular in Africa.
- Solution: Examine queries and rerank based on language-based collection weights.

Bantu Language IR

Nyasha Katemauswa, Michael Kyeyune, (U/G); Nkosana Malumba, Katlego Moukangwe, Sinead Urisohn, Andre Lopes, Bsc(Hons); Catherine Chavula, Joseph Telemala (PhD - current)

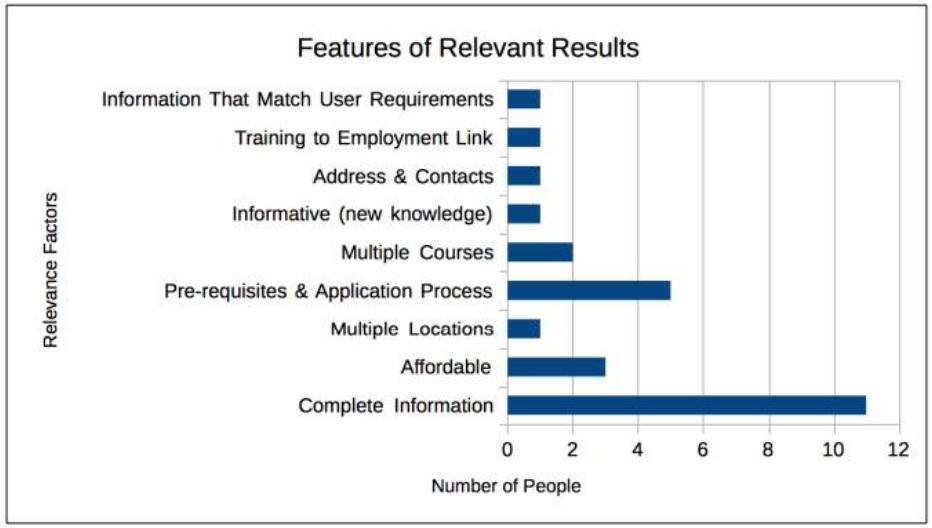
- Search engines in Bantu languages in South Africa/Malawi/Tanzania/etc.
- Many core IR algorithms are unchanged but some language-specific algorithms needed: Language identification; Text pre-processing and normalization; Ranking and reranking...

Studies: IR/DM for Development

IR for Development

Gina Paihama, PhD (current)

- How can we give users directed results to address unemployment?
- Relevance is more specific here:



DM for Development

Selvas Mwanza, PhD (current)

- Can we use Twitter data to evaluate developmental measures in society (e.g., level of free speech)?
  - We have found an association between what people discuss (politics vs. entertainment) and how.

class	t_n	n	m	ϕ	p	C	r	$r^{\rightleftharpoons}$	l_G	Q	type
P-S_p	857	394	423	3	0.005463634	0.0005309359	0.004471068	0	1.09636	0.5290786	forward
S_p	805	582	567	5	0.003353621	0.0007443553	-0.01618863	0.003527337	1.348158	0.7023055	combined
E	3476	447	302	2	0.003029665	0.01648352	0.06462597	0.006622517	1.103858	0.9404357	mention
P	456	140	118	2	0.01212744	0	-0.05746937	0	1.008403	0.8601336	forward
S_p	12631	1445	1707	9	0.00163617	0.009588494	-0.007949564	0.01405975	2.964084	0.8148845	combined
E-S_p	1736	427	431	3	0.004738815	0.001455604	-0.004947816	0.004640371	1.103858	0.8068594	mention

Studies: IR in Low Resource Environments

Bantu Language IR: Speech UI

Morebodi Modise, MSc

- Speech-driven mobile search interface in isiXhosa.
- Works well, but educated people want English!

(a) Query submission interface

(b) Detecting voice queries

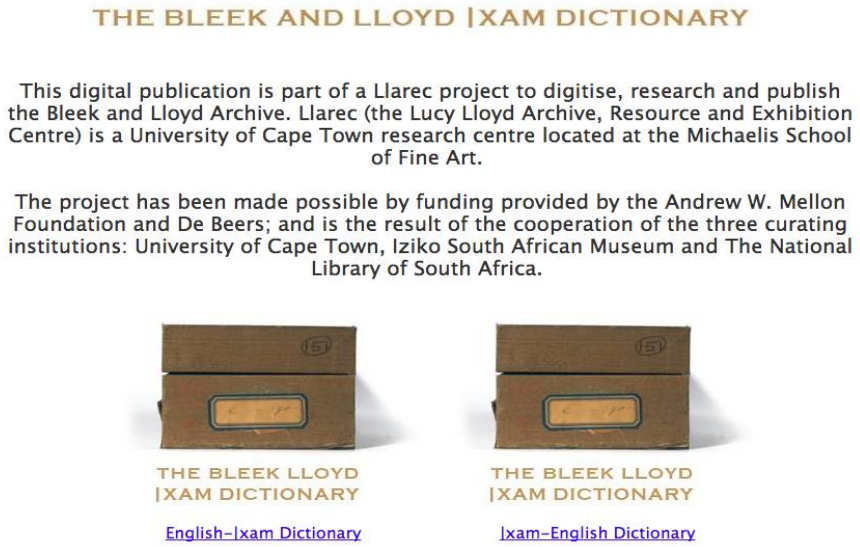
(c) Detected voice query with list of voice results

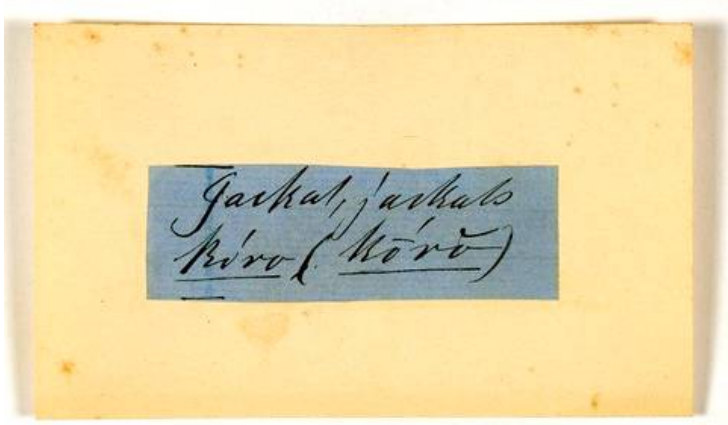
Fig. 3: Mobile voice interface

Bleek and Lloyd: Dictionary

Lebogang Molwantoa, Sanvir Manilal, Kyle Williams, BSc(Hons)

- Visual dictionary – pictures of words.
- Find meanings of words in stories by image search.





Bleek and Lloyd: Transcription

Kyle Williams, MSc; Ngoni Munyaradzi, MSc

- Using machine learning to transcribe |Xam.
- Training data manually generated.
- 45% accuracy at best.

- Crowdsourcing had 10% better performance.
  - Answer determined by agreement among 3 amateur transcribers.

Studies: Low-Resource Repositories

Repositories in Low-Resource Environments

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

  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

What is the net effect?

What is a Digital Object Repository?

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

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?

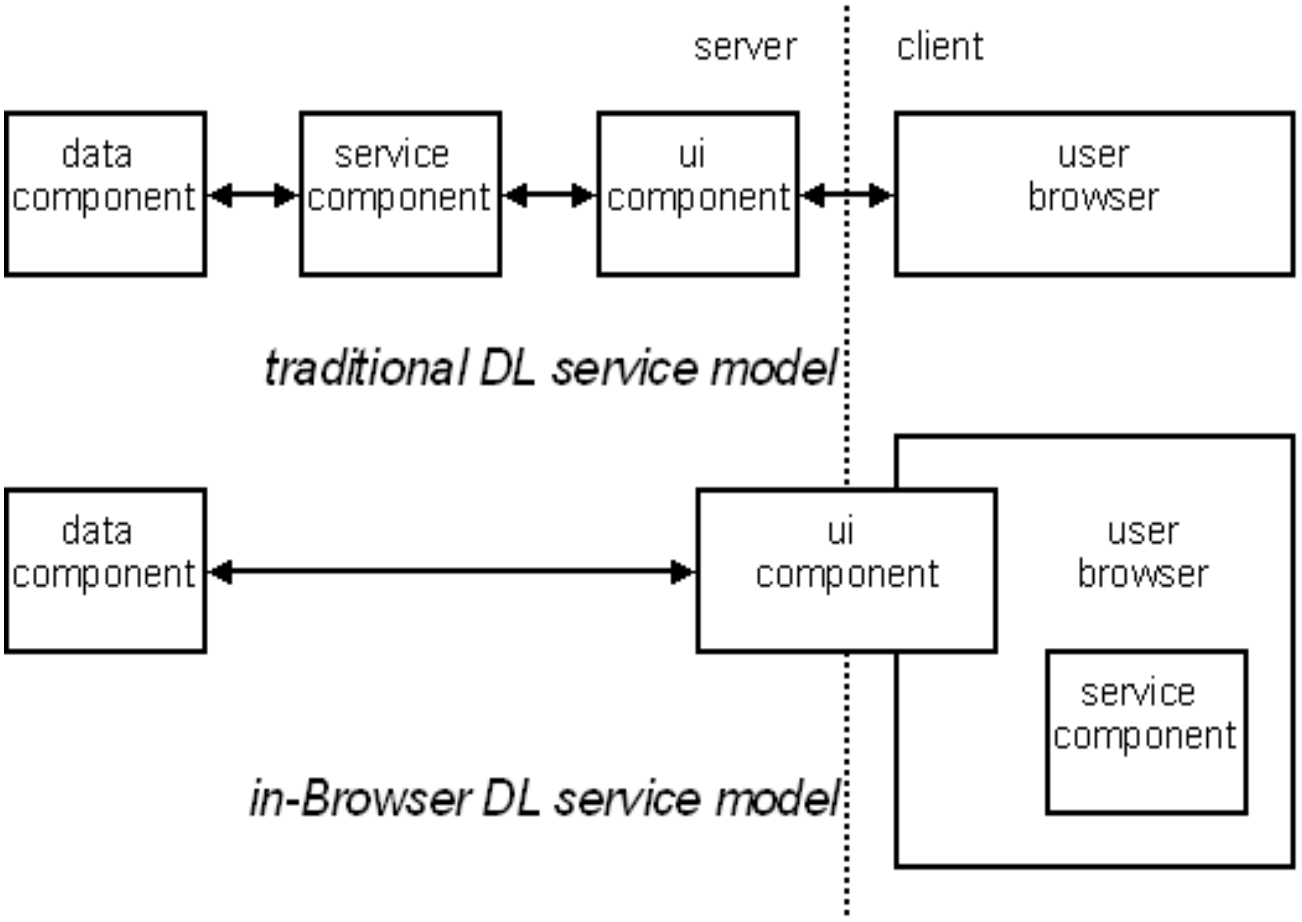
  OR

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

Approach 1: XML + XSLT → XHTML

- Encode all descriptive information using XML files for objects, users, etc.
- Use XSL transformations to convert the XML metadata into corresponding HTML page views.
- Also, pre-process XML to create search/browse indices.

Approach 2: in-Browser Services



Approach 3: Selected Online Services

- Use online Web applications to add items, comments, annotation, etc.

- Web applications must modify XML and regenerate portions of static site.
  - No dynamic page generation.

FHYA Prototype 1/3

500 FIVE HUNDRED YEAR ARCHIVE

Search FHYA Collections

Enter your search terms:

GO

Browse FHYA Collections

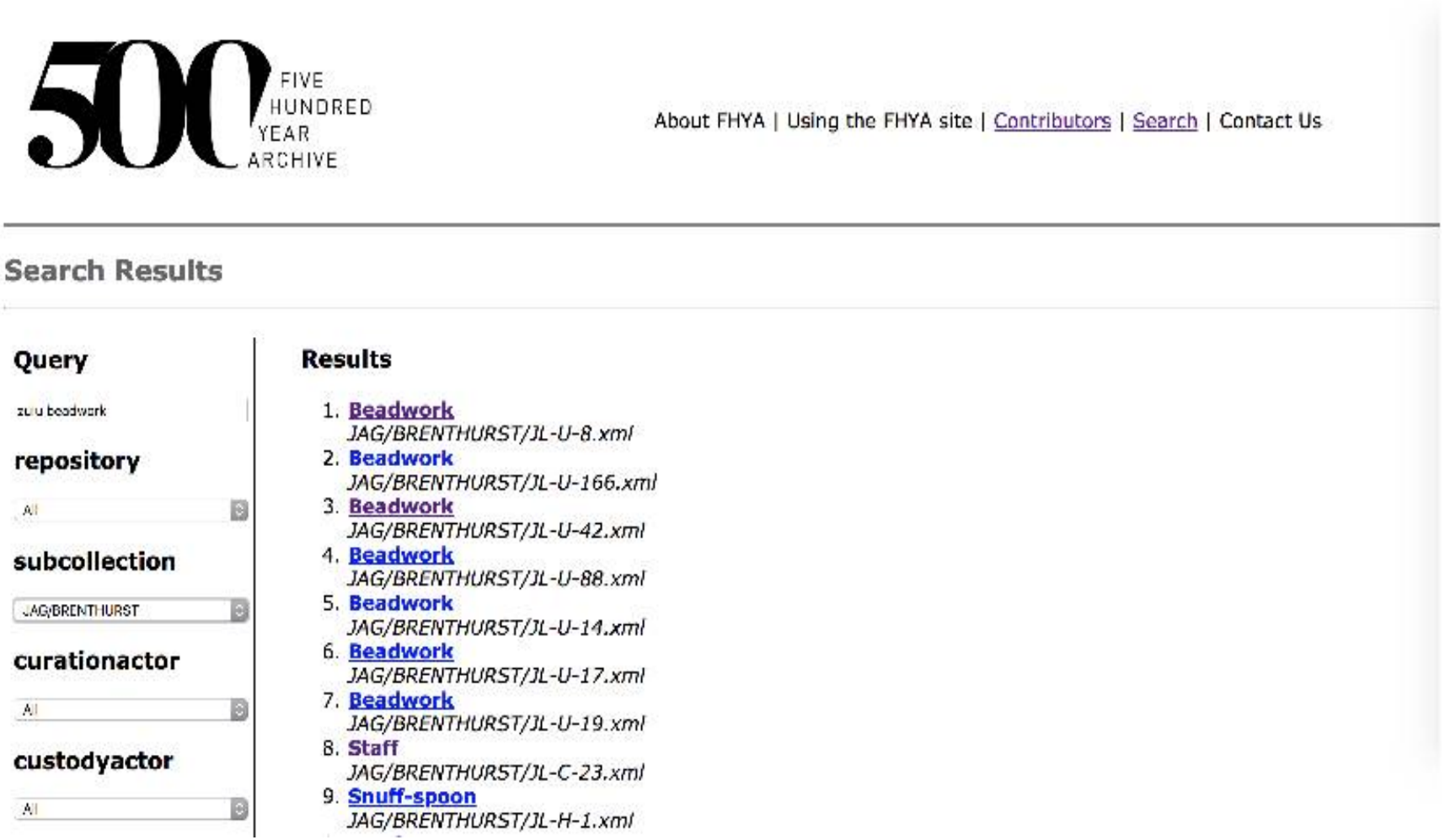
KwaZulu-Natal MUSEUM
an agency of the Department of Arts and Culture

KZN Museums

Wits University

Public Contributions

FHYA Prototype 3/3



Complex Query Results: Ave. Perform

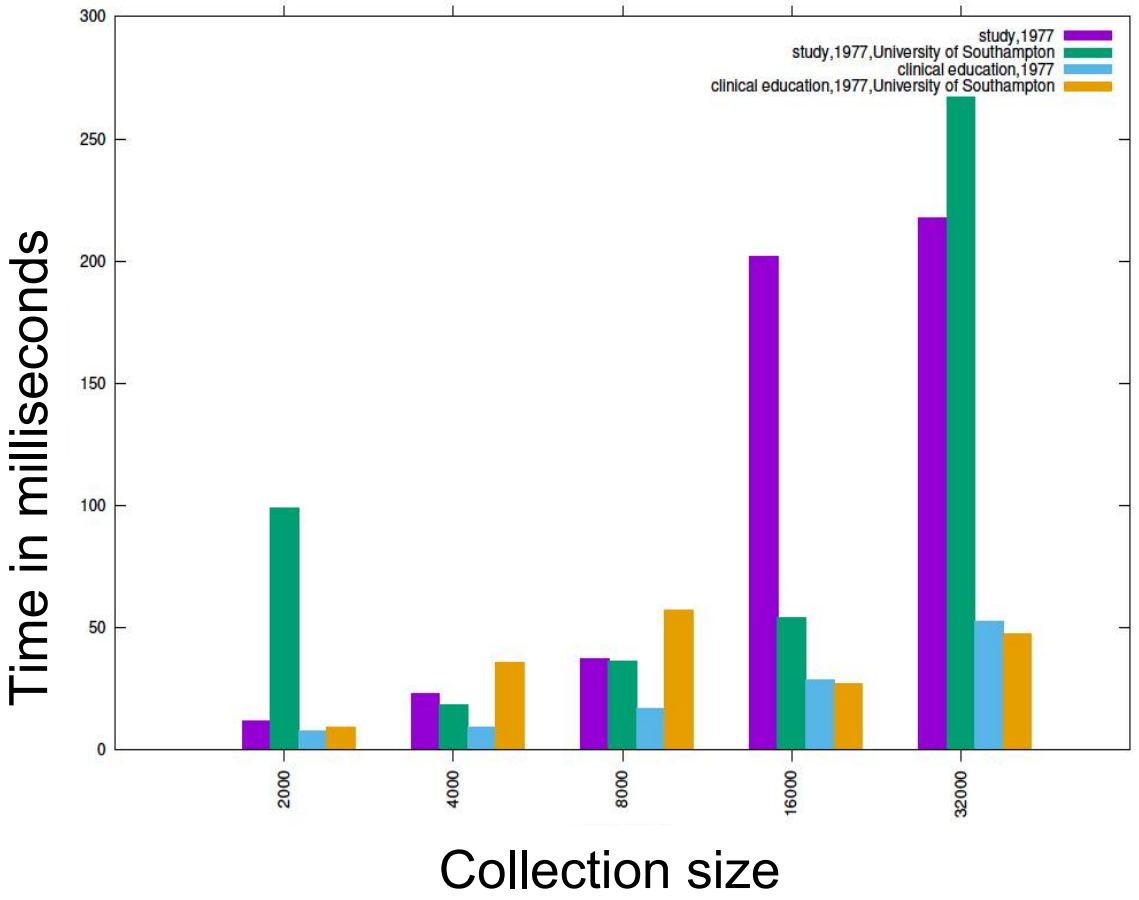


Fig. 5. Average times for faceted search/browse queries of difference complexities across all collection sizes.

What next?

Where we are

- Some early successes but still:
  - Many languages, with
  - Few documents,
  - Few resources (money/users), and
  - Lots of mixing of languages in queries and documents.
- Lots of work still needed
- Lots of opportunities for research

What should we do

- Think differently about data.
- Think differently about IR/DM.
  - Some ideas are universal (indexing?).
  - Some ideas are not (online, text, Unicode).
- Invent completely new solutions.
- Change the world!

Why AFIRM?

- Boost research in IR/DM in Africa.
- Reconnect African researchers with international community.
- Build a community of African researchers.
- Things to ponder:
  - AFIRM 2021 – where?
  - How can AFIRM improve the model?
  - What else should we do as a community?
  - Can we set a common challenge as the basis for research? What challenge?

questions, comments, ...

More information, papers, etc. on
http://dl.cs.uct.ac.za/

enkosi
hamba kakuhle
thank you and go well