Web Archive Workshop

Finding and Analyzing Web Archives

Ian Milligan
Assistant Professor

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

Links/resources at https://github.com/ianmilligan1/iSchool-Workshop

Workshop Plan

A. Web Archive Collections (Nicholas Worby)

B. Web Archive Analysis (Ian Milligan)

    A. Where to find them?

    B. How to analyze them?

    C. Hands-on with warcbase and network analysis

How to use these slides?

- If you're in the room - awesome! They're available for download, so you don't have to scrawl down links.
- If you're not in the room - too bad! But that's OK. I'm using these as a seed list of links to walk people through. But you should still be able to follow along too.

Where to find web archives?

Collect them yourself?

- Archive-It Subscription
- wget-based web archive
  - wget “http://www.ianmilligan.ca/“ --warc-file="im"
  - https://github.com/ruebot/arxivdaleascii
  - http://freedaleaskey.plggta.org/
- WebRecorder: https://webrecorder.io/

Internet Archive

- https://archive.org/details/wide00002
- https://webarchive.jira.com/wiki/display/ARS/Archive-It+Research+Services (i.e. ask your Archive-It partners if you're interested)

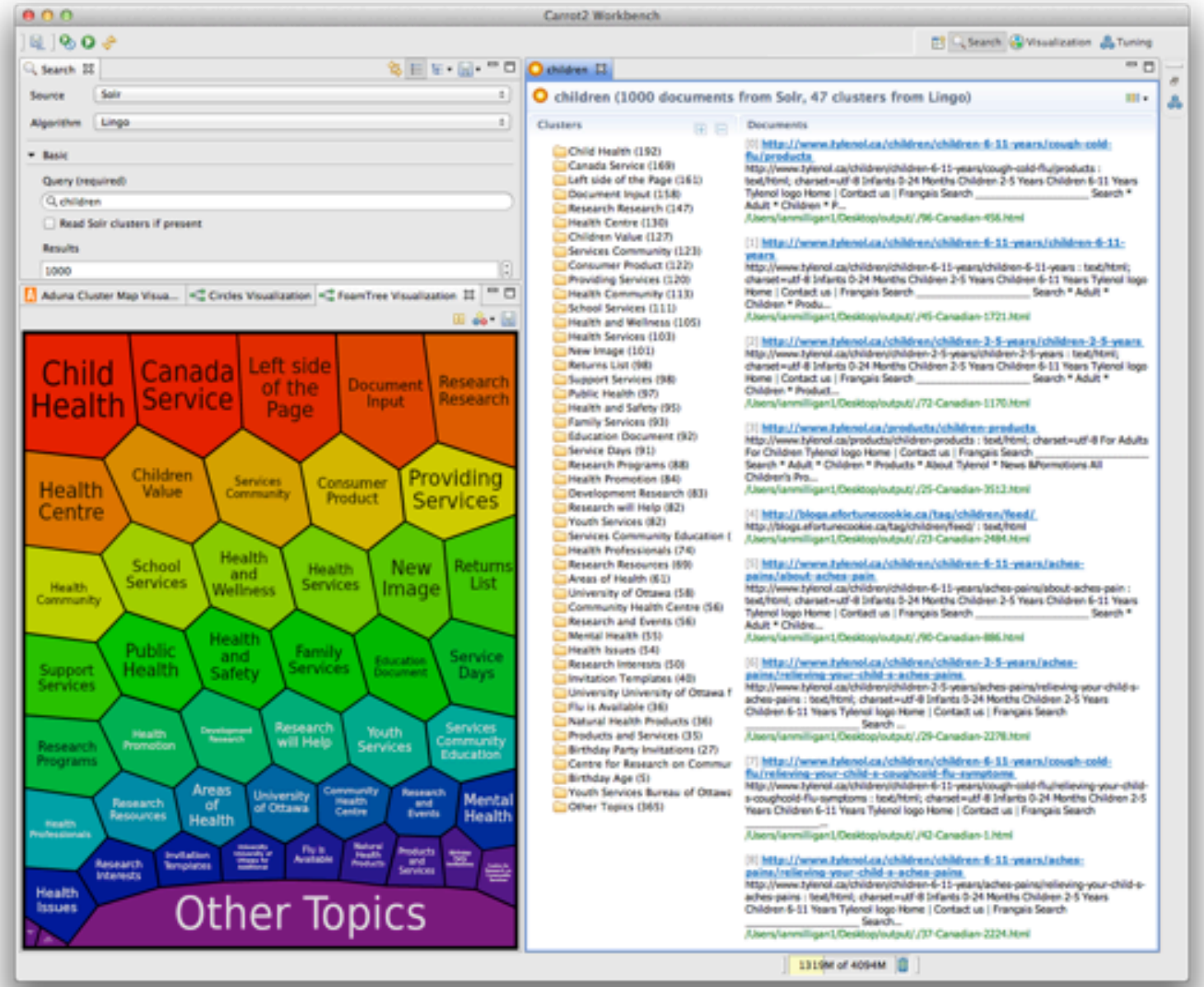
Other Sources

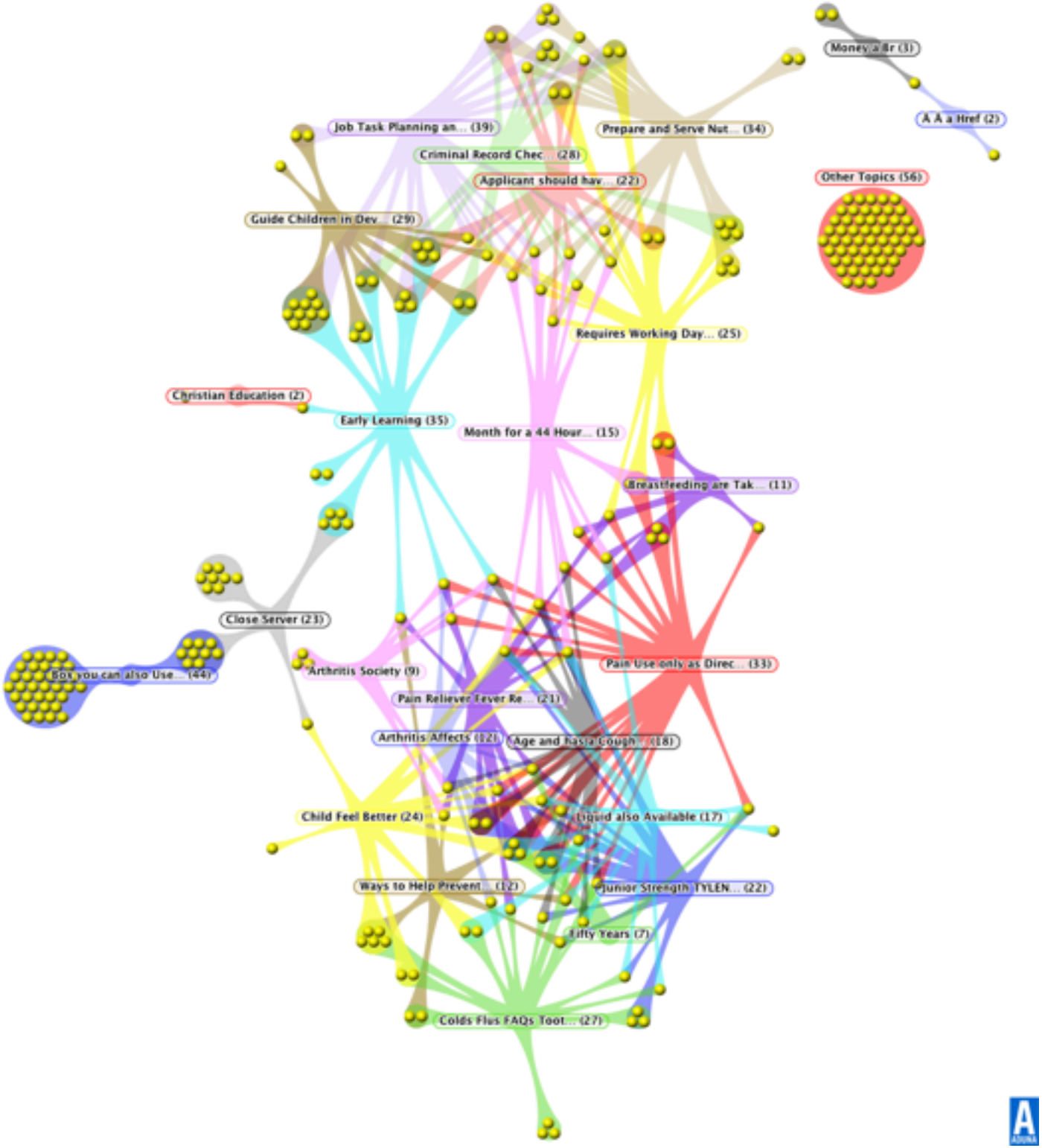
- Government of Canada Web Archive
- .gov Web Archive (end of congressional crawls)
- Custom-generated ones from collecting institutions
- **Come to our web archive hackathon? (3 - 5 March 2016)**

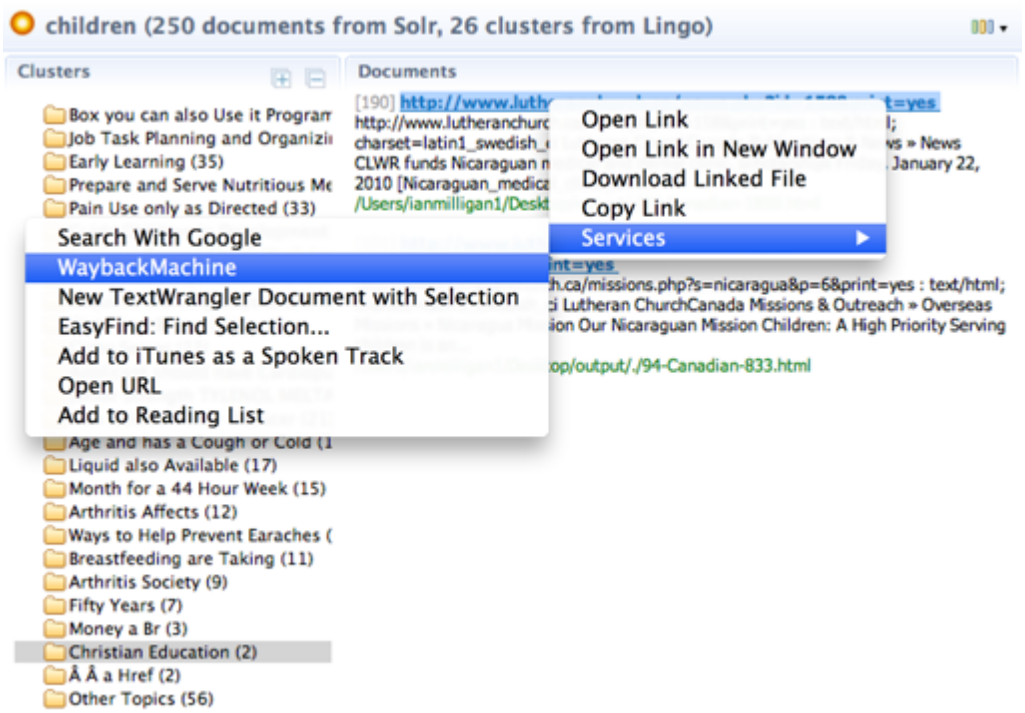
How to Analyze them?

- What is Solr?
- UKWA WARC Indexer: https://github.com/ukwa/webarchive-discovery/wiki/Quick-Start or https://github.com/ukwa/webarchive-discovery/tree/master/warc-indexer
- Hadoop Indexer: https://github.com/ukwa/webarchive-discovery/tree/master/warc-hadoop-indexer
- Speed issues

Basic keyword searching


Carrot2 Workbench
Search
Visualization
Tuning
Source
Solr
Algorithm
Lingo
Basic
Query (required)
children
Read Solr clusters if present
Results
1000
Aduna Cluster Map Visua...
Circles Visualization
FoamTree Visualization
Child Health
Canada Service
Left side of the Page
Document Input
Research Research
Health Centre
Children Value
Services Community
Consumer Product
Providing Services
Health Community
School Services
Health and Wellness
Health Services
New Image
Returns List
Public Health
Health and Safety
Family Services
Education Document
Service Days
Support Services
Research Programs
Health Promotion
Development Research
Research will Help
Youth Services
Services Community Education
Areas of Health
University of Ottawa
Community Health Centre
Research and Events
Mental Health
Research Resources
Health Professionals
Research Interests
Invitation Templates
Flu is Available
Natural Health Products
Products and Services
Health Issues
Other Topics
children
children (1000 documents from Solr, 47 clusters from Lingo)
Clusters
Child Health (192)
Canada Service (169)
Left side of the Page (161)
Document Input (158)
Research Research (147)
Health Centre (130)
Children Value (127)
Services Community (123)
Consumer Product (122)
Providing Services (120)
Health Community (113)
School Services (111)
Health and Wellness (105)
Health Services (103)
New Image (101)
Returns List (98)
Support Services (98)
Public Health (97)
Health and Safety (95)
Family Services (93)
Education Document (92)
Service Days (91)
Research Programs (88)
Health Promotion (84)
Development Research (83)
Research will Help (82)
Youth Services (82)
Services Community Education (
Health Professionals (74)
Research Resources (69)
Areas of Health (61)
University of Ottawa (58)
Community Health Centre (56)
Research and Events (56)
Mental Health (55)
Health Issues (54)
Research Interests (50)
Invitation Templates (40)
University University of Ottawa f
Flu is Available (36)
Natural Health Products (36)
Products and Services (35)
Birthday Party Invitations (27)
Centre for Research on Commur
Birthday Age (5)
Youth Services Bureau of Ottawa
Other Topics (365)
Documents
[0] http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/products
http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/products : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____ Search * Adult * Children * P...
/Users/ianmilligan1/Desktop/output/./96-Canadian-456.html
[1] http://www.tylenol.ca/children/children-6-11-years/children-6-11-years
http://www.tylenol.ca/children/children-6-11-years/children-6-11-years : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____ Search * Adult * Children * Produ...
/Users/ianmilligan1/Desktop/output/./45-Canadian-1721.html
[2] http://www.tylenol.ca/children/children-2-5-years/children-2-5-years
http://www.tylenol.ca/children/children-2-5-years/children-2-5-years : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____ Search * Adult * Children * Product...
/Users/ianmilligan1/Desktop/output/./72-Canadian-1170.html
[3] http://www.tylenol.ca/products/children-products
http://www.tylenol.ca/products/children-products : text/html; charset=utf-8 For Adults For Children Tylenol logo Home | Contact us | Français Search _____ Search * Adult * Children * Products * About Tylenol * News &Promotions All Children's Pro...
/Users/ianmilligan1/Desktop/output/./25-Canadian-3512.html
[4] http://blogs.efortunecookie.ca/tag/children/feed/
http://blogs.efortunecookie.ca/tag/children/feed/ : text/html
/Users/ianmilligan1/Desktop/output/./23-Canadian-2484.html
[5] http://www.tylenol.ca/children/children-6-11-years/aches-pains/about-aches-pain
http://www.tylenol.ca/children/children-6-11-years/aches-pains/about-aches-pain : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____ Search * Adult * Childre...
/Users/ianmilligan1/Desktop/output/./90-Canadian-886.html
[6] http://www.tylenol.ca/children/children-2-5-years/aches-pains/relieving-your-child-s-aches-pains
http://www.tylenol.ca/children/children-2-5-years/aches-pains/relieving-your-child-s-aches-pains : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____ Search ...
/Users/ianmilligan1/Desktop/output/./29-Canadian-2278.html
[7] http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-coughcold-flu-symptoms
http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-coughcold-flu-symptoms : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____...
/Users/ianmilligan1/Desktop/output/./42-Canadian-1.html
[8] http://www.tylenol.ca/children/children-6-11-years/aches-pains/relieving-your-child-s-aches-pains
http://www.tylenol.ca/children/children-6-11-years/aches-pains/relieving-your-child-s-aches-pains : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Français Search _____ Search...
/Users/ianmilligan1/Desktop/output/./37-Canadian-2224.html
1319M of 4094M


Money a Br (3)
A A a Href (2)
Other Topics (56)
Job Task Planning an... (39)
Prepare and Serve Nut... (34)
Criminal Record Chec... (28)
Applicant should hav... (22)
Guide Children in Dev... (29)
Requires Working Day... (25)
Christian Education (2)
Early Learning (35)
Month for a 44 Hour... (15)
Breastfeeding are Tak... (11)
Close Server (23)
Bot you can also Use... (44)
Arthritis Society (9)
Pain Use only as Direc... (33)
Pain Reliever Fever Re... (21)
Arthritis Affects (12)
Age and hasta Cough ... (18)
Child Feel Better (24)
Liquid also Available (17)
Ways to Help Prevent... (12)
Junior Strength TYLEN... (22)
Fifty Years (7)
Colds Flus FAQs Toot... (27)


children (250 documents from Solr, 26 clusters from Lingo)
Clusters
Documents
Box you can also Use it Program
Job Task Planning and Organizi
Early Learning (35)
Prepare and Serve Nutritious Me
Pain Use only as Directed (33)
Age and has a Cough or Cold (1
Liquid also Available (17)
Month for a 44 Hour Week (15)
Arthritis Affects (12)
Ways to Help Prevent Earaches (
Breastfeeding are Taking (11)
Arthritis Society (9)
Fifty Years (7)
Money a Br (3)
Christian Education (2)
Â Â a Href (2)
Other Topics (56)
[190] http://www.luth
http://www.lutheranchurc
charset=latin1_swedish_
CLWR funds Nicaraguan n
2010 [Nicaraguan_medica
/Users/ianmilligan1/Deskt
Open Link
Open Link in New Window
Download Linked File
Copy Link
Services
Search With Google
WaybackMachine
New TextWrangler Document with Selection
EasyFind: Find Selection...
Add to iTunes as a Spoken Track
Open URL
Add to Reading List
January 22,
h.ca/missions.php?s=nicaragua&p=6&print=yes : text/html;
Lutheran ChurchCanada Missions & Outreach » Overseas
ion Our Nicaraguan Mission Children: A High Priority Serving
op/output/./94-Canadian-833.html


Lutheran Church-Canada
web.archive.org/web/20110714130258/http://www.lutheranchurc...
http://www.lutheranchurch.ca/news.php?id=158&print=yes
Go
3 captures
DEC
JUL
14
2010
2011
LUTHERAN CHURCH–CANADA
ÉGLISE LUTHÉRIENNE du CANADA
CLWR funds Nicaraguan medical and dental clinic, scholarships
Friday, January 22, 2010
WINNIPEG – Canadian Lutheran World Relief (CLWR) has announced $36,500 in funding for two Lutheran Church–Canada (LCC) programs in Nicaragua this year.
The announcement was made as Iglesia Luterana Sinodo de Nicaragua (ILSN) prepares for its first biennial convention and includes new money for a medical and dental clinic and increased school scholarships.
The medical clinic, which began operations in May 2009, is open every Thursday beginning at 8 a.m. and remains open until all patients have been seen.
The clinic is staffed by a doctor and a dentist, who see an average of 40-45 patients each week, and provides common medications because many patients are too poor to purchase them.
A medical clinic in Nicaragua.
CLWR will continue supporting the Christian Children Education Program.
The program, conducted in all 23 congregations of ILSN, provides an average of 25 scholarships in each community to the neediest children.
The scholarships include the required school uniforms, shoes, backpacks and school supplies.
Each child is also enrolled in the tutoring and Christian-education class held five days a week when children are not in school (Children after school in the morning or in the afternoon.)
These classes, held in the churches and led by teachers and deaconesses, provide tutoring and homework support for the children in math, Spanish and other subjects. A portion of the time is also set aside for Christian education and cultural activities.
More than 750 children are enrolled in the program. CLWR has provided support for about 250 children.
Since 1999, CLWR has partnered with LCC to support community-development projects.
Robert Granke, executive director of CLWR, visited congregations of the ILSN in November. You can read more about his visit at www.
www.lccontheroad.ca , The Canadian Lutheran or in the forthcoming issue of CLWR's Partnership newsletter due out in early February.

UKWA Shine

- https://github.com/ukwa/webarchive-discovery/wiki
- https://github.com/ianmilligan1/WAHR/blob/master/walkthroughs/WebArchive-Discovery-and-Shine-Doc.md
- WebArchives.ca

Web Archive Analysis with Warcbase

Warcbase

- Jimmy Lin (University of Waterloo)
- CS-History collaboration - they have the expertise, we have the questions!

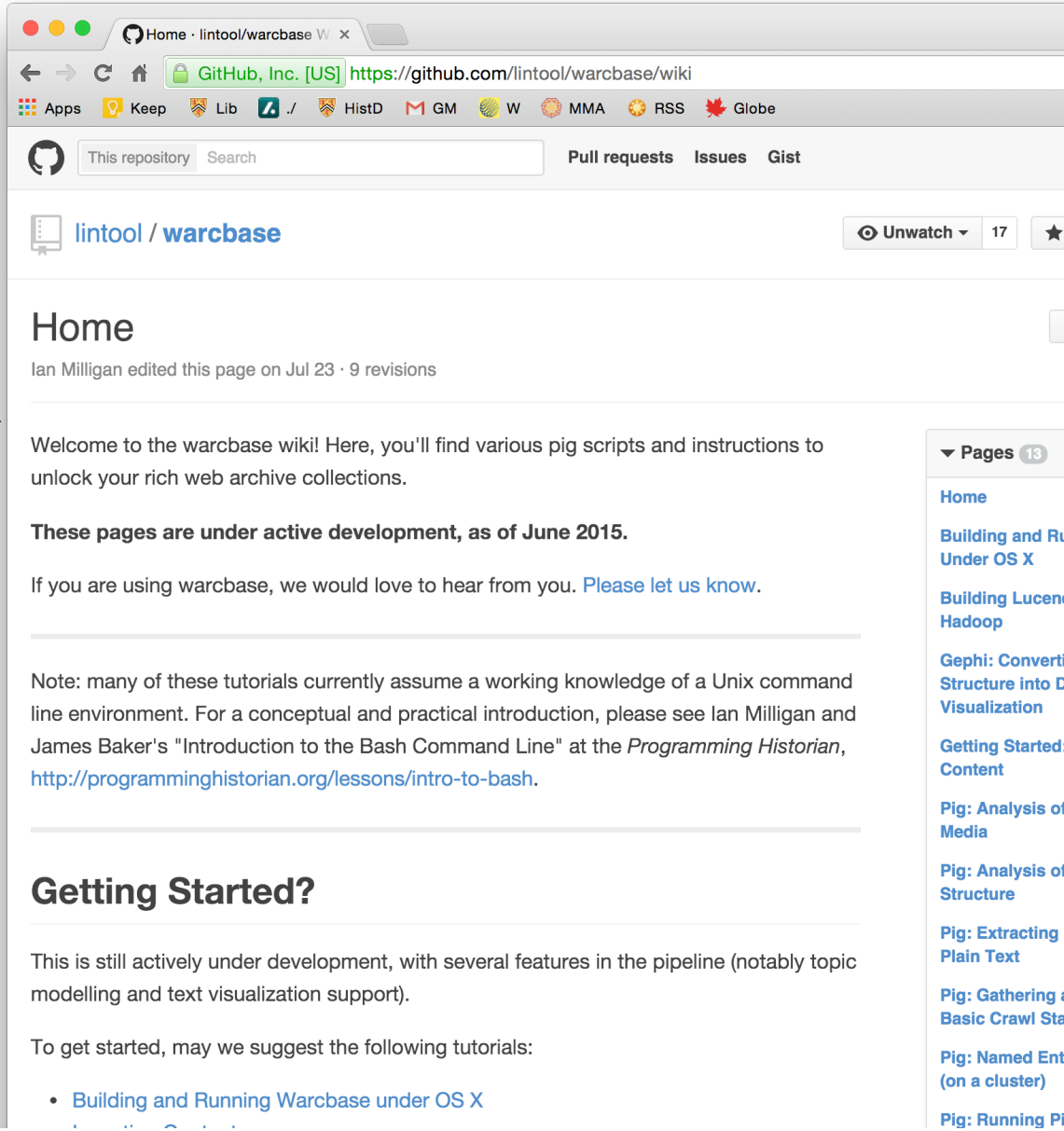
What is warcbase?

- Warcbase is a web archive platform, not a single program. Its capabilities comprise two main categories:
  - Analysis of web archives using the Pig or Spark programming languages, and assorted helper scripts and utilities
  - Web archive database management, with support for the HBase distributed data store, and OpenWayback integration providing a friendly web interface to view stored websites
- One can take advantage of the analysis tools (1) without bothering with the database management aspect of Warcbase -- in fact, most digital humanities researchers will probably find the former more useful.

warcbase documentation

- https://github.com/lintool/warcbase/wiki

- Walkthrough of “Spark Extracting Domain Level Plain Text”, “Pig: Analysis of Links to Social Media,” and “Pig: Analysis of Site Link Structure.”



Text Analysis

- Date-ordered plain-text
- Once you have it:
  - **Voyant Tools** if small (http://voyant-tools.org/)
  - R/Mathematica/Python if larger
  - PySpark?
  - Back into **solr**?

Text Analysis

- Topic modelling? (http://mallet.cs.umass.edu/)
- Stanford NER Visualization (http://ianmilligan1.github.io/Ner-Viz/?csv=greenparty.csv)

Network Analysis

- https://github.com/lintool/warcbase/wiki/Gephi:-Converting-Site-Link-Structure-into-Dynamic-Visualization
- Considering moving to d3.js?
- But right now, Gephi: http://gephi.github.io/

Network Analysis

- Gephi
  - Requires Java JDK 1.7 (Java 1.8 breaks it)
  - Debugging on all systems is a PIA
- If you can run it on your own system, that's best
- If not, you can download a VM and install it there: http://ianmilligan.ca/historycrawler/

Follow along - sample dataset at https://github.com/ianmilligan1/iSchool-Workshop

political-links.gdf

Hands-on portion/
Questions and Answers,
etc.

Funding Acknowledgements

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario
MINISTRY OF RESEARCH & INNOVATION

UNIVERSITY OF WATERLOO