

'An Infinite Archive?'
**Historical Explorations in
the Internet Archive's Wide
Web Scrape**

Ian Milligan, PhD
Assistant Professor of History
i2millig@uwaterloo.ca



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



[http://en.wikipedia.org/wiki/
File:Internet_map_1024.jpg](http://en.wikipedia.org/wiki/File:Internet_map_1024.jpg)

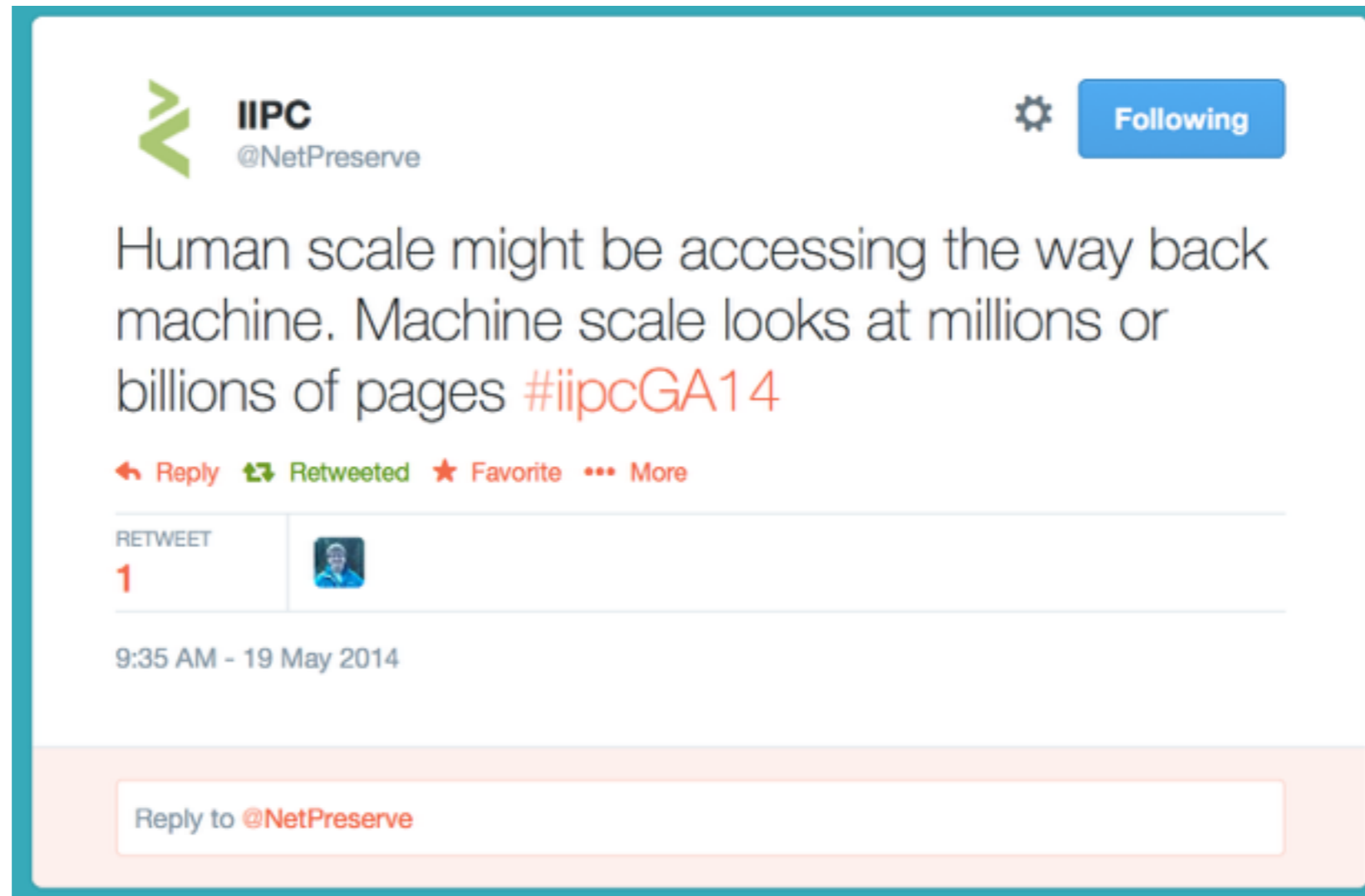
Why?

Historians need to think
about **Computational
Methods** in an era of
web archives.

“.... [n]ow expectations have inverted. Everything may be recorded and preserved*, at least potentially.”

- James Gleick, *The Information*

** an overstatement, of course, but a useful one*



We have too much information to make sense of with normal methods.

The 80TB Wide Web Scrape

[March - December 2011]

ca,yorku,justlabour)/ 20110714073726
http://www.justlabour.yorku.ca/ text/html
302 3I42H3S6NNFQ2MSVX7XZKYAYSCX5QBYJ
http://www.justlabour.yorku.ca/index.php?
page=toc&volume=16 - 462 880654831
WIDE-20110714062831-crawl416/
WIDE-20110714070859-02373.warc.gz

Top-Level Domain	Number of Distinct URLs Downloaded in Sample	Number of Overall URLs in Wide Web Scrape (selected domains)	Percentage of URLs Captured
.com	29,219,706	1,260,409,874	2.32%
.org	2,489,050	96,681,268	2.57%
.net	2,438,903	140,726,805	1.73%
.edu	350,482	6,620,283	5.29%
.gov	97,484	2,205,332	4.42%
.mil	10,268	103,507	9.92%
.ca	622,365	8,512,275	7.31%
.uk	464,991	21,870,821	2.13%
.fr	239,160	13,654,404	1.75%
.in	105,287	3,736,316	2.82%
.cn	5,499,593	133,105,864	4.13%
.ke	4883	37,871	12.89%
TOTAL	41,542,172	1,687,664,620	2.46%

Methods

(or the fun of playing with
WARC files themselves)

#<<< Ian Milligan | A Digital, Public, and Youth Historian of 20th-Century Canada (p1 of 7)
#Ian Milligan Feed Ian Milligan Comments Feed Ian Milligan
WordPress.com

Ian Milligan

A Digital, Public, and Youth Historian of 20th-Century Canada

Menu

Skip to content

- * Home
- * About Me
- * CV
- * Teaching
- * My Digital History Work
 - + Current Work/Project Description
 - + Project Proposal (SSHRC)
- * "Rebel Youth"

Visualizing Locations in the Internet Archive .ca Wide Scrape Sample

Standard

Taking the full text of my sample of **my Canadian (.ca only) websites** (currently being finessed to amount 622,365 URLs out of a scrape total of 8,512,275 or 7.31%), I ran it through **Stanford NER** and extracted popular locations, organizations, and people. This is a morning's work, mainly as I let my desktop crunch away at some other stuff, so I really need to preface the post that the data has not been cleaned up.

The results were interesting but fairly dry: "Canada" was the top location, for example, followed by Ontario, Toronto, Ottawa, Alberta, etc. The United States comes out as under-represented mainly because we have so many spellings of the word (US, u s, America, United States, etc.). There will be a similar issue with the United Kingdom. If this turns into 'real' research rather than tinkering, again, there's a lot of cleaning up to do. But overall, we can get a rough sense of different countries and how they appeared in this sample.

Thanks to **IBM Many Eyes** we can throw this stuff at the wall and see what comes out.

Screen Shot 2014-02-06 at 11.47.16 AM

In this graphic, we see different countries and how they are mentioned. With the caveats that this is rough data, we can see the big parties

-- press space for next page --

Arrow keys: Up and Down to move. Right to follow a link; Left to go back.
H)elp O)ptions P)rint G)o M)ain screen Q)uit /=search [delete]=history list



[Canada faces US in battle for women's hockey Olympic gold](#)

[The Globe and Mail](#) - 3 hours ago

They started out together on the 1998 **Canadian** Olympic team. Jayna Hefford and Hayley Wickenheiser – the last two players remaining from ...

[Canada vs. USA: Olympic women's hockey final live coverage](#)

[Canada.com](#) - 11 minutes ago

[Canadians Just Barely Survive A Historic Hockey Scare](#)

[Toronto Star](#) - Feb 19, 2014



[CTV News](#)

[Globalnews.ca](#)

[Bleacher Report](#)

[CBSSports.com](#)

[all 2840 news sources »](#)



Carrot2 Workbench

Search Visualization Tuning

Search []

Source Solr

Algorithm Lingo

Basic

Query (required)

children

Read Solr clusters if present

Results

1000

Aduna Cluster Map Visual... Circles Visualization FoamTree Visualization

children (1000 documents from Solr, 47 clusters from Lingo)

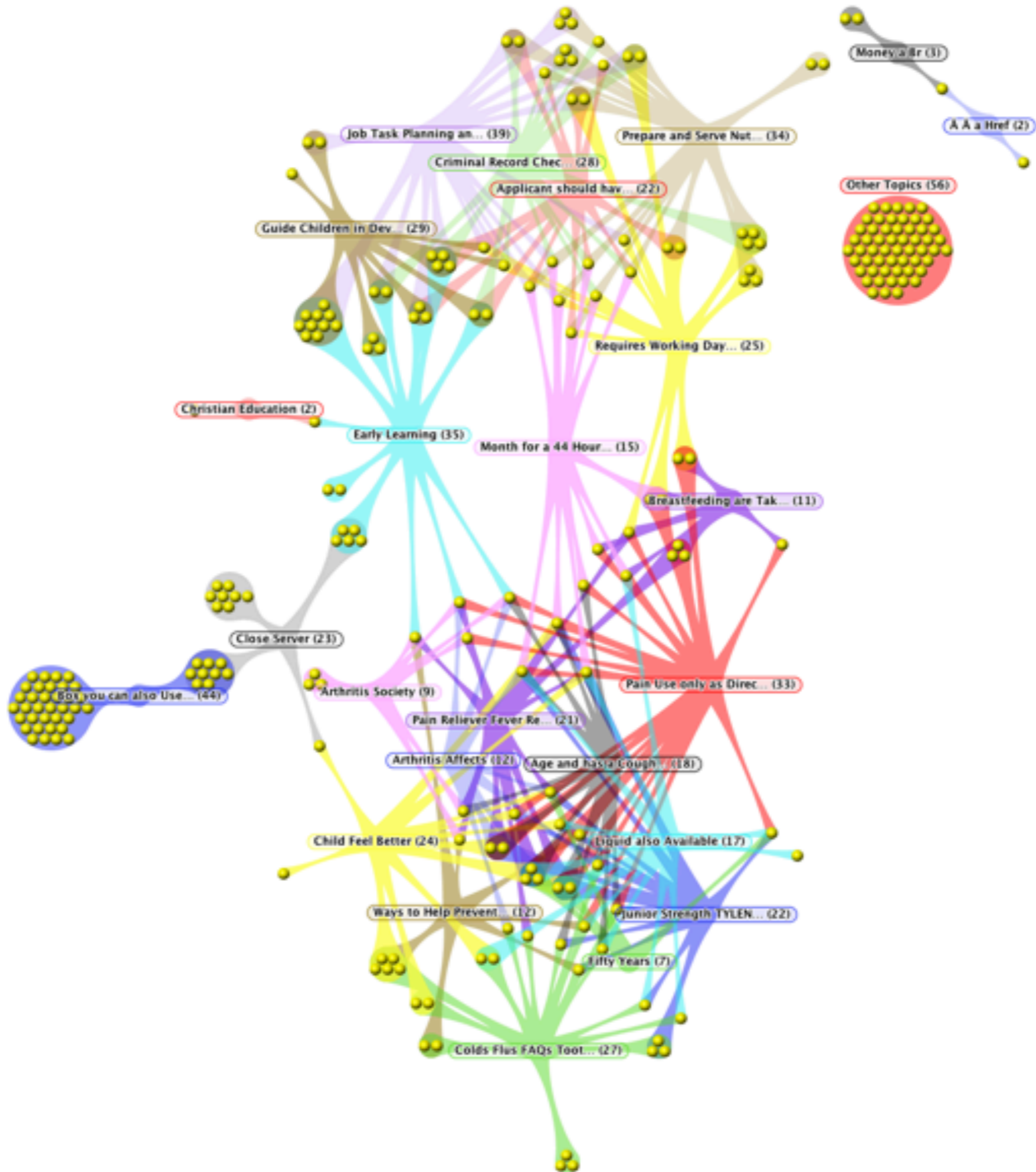
Clusters

- Child Health (192)
- Canada Service (169)
- Left side of the Page (161)
- Document Input (158)
- Research Research (147)
- Health Centre (136)
- Children Value (127)
- Services Community (123)
- Consumer Product (120)
- 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 (74)
- Health Professionals (74)
- Research Resources (69)
- Areas of Health (63)
- University of Ottawa (58)
- Community Health Centre (56)
- Research and Events (56)
- Mental Health (51)
- Health Issues (54)
- Research Interests (50)
- Invitation Templates (48)
- University of Ottawa (47)
- Flu is Available (38)
- Natural Health Products (36)
- Products and Services (35)
- Birthday Party Invitations (27)
- Centre for Research on Commun (27)
- Birthday Age (5)
- Youth Services Bureau of Ottawa (5)
- Other Topics (365)

Documents

- <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 | Francais Search Adult * Children * P...
/Users/annmulligan/Desktop/output/76-Canadian-496.html
- <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 | Francais Search Adult * Children * Produ...
/Users/annmulligan/Desktop/output/76-Canadian-1721.html
- <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 | Francais Search Search * Adult * Children * Products * About Tylenol * News & Promotions All Children's Pro...
/Users/annmulligan/Desktop/output/25-Canadian-3512.html
- <http://blogs.afortunecookie.ca/tag/children/feed/>
http://blogs.afortunecookie.ca/tag/children/feed/ : text/html
/Users/annmulligan/Desktop/output/73-Canadian-2494.html
- <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 | Francais Search Adult * Childre...
/Users/annmulligan/Desktop/output/76-Canadian-886.html
- <http://www.tylenol.ca/children/children-2-5-years/aches-pains/relieving-your-child-s-aches-pain>
http://www.tylenol.ca/children/children-2-5-years/aches-pains/relieving-your-child-s-aches-pain : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search Search ...
/Users/annmulligan/Desktop/output/29-Canadian-2278.html
- <http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-cough-cold-flu-symptoms>
http://www.tylenol.ca/children/children-6-11-years/cough-cold-flu/relieving-your-child-s-cough-cold-flu-symptoms : text/html; charset=utf-8 Infants 0-24 Months Children 2-5 Years Children 6-11 Years Tylenol logo Home | Contact us | Francais Search Search...
/Users/annmulligan/Desktop/output/77-Canadian-2224.html

1319M of 4094M



children (250 documents from Solr, 26 clusters from Lingo)

Clusters

- Box you can also Use it Program
- Job Task Planning and Organiz
- Early Learning (35)
- Prepare and Serve Nutritious Me
- Pain Use only as Directed (33)
- 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
- 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)
- À À a Href (2)
- Other Topics (56)

Documents

[190] <http://www.luth...> <http://www.lutheranchurc...> <http://www.lutheranchurc...>

CLWR funds Nicaraguan n
2010 [Nicaraguan_medice
/Users/ianmilligan1/Desk

Open Link
Open Link in New Window
Download Linked File
Copy Link
Services

nt=yes
h.ca/missions.php?s=nicaragua&p=6&print=yes : text/html;
Lutheran ChurchCanada Missions & Outreach » Overseas
Our Nicaraguan Mission Children: A High Priority Serving
op/output/,/94-Canadian-833.html

Lutheran Church-Canada x

web.archive.org/web/20110714130258/http://www.lutheranchurc...

INTERNET ARCHIVE WaybackMachine 3 captures 5 Dec 10 - 14 Jul 11

DEC 14 2010 JUL 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.

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 attend 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.

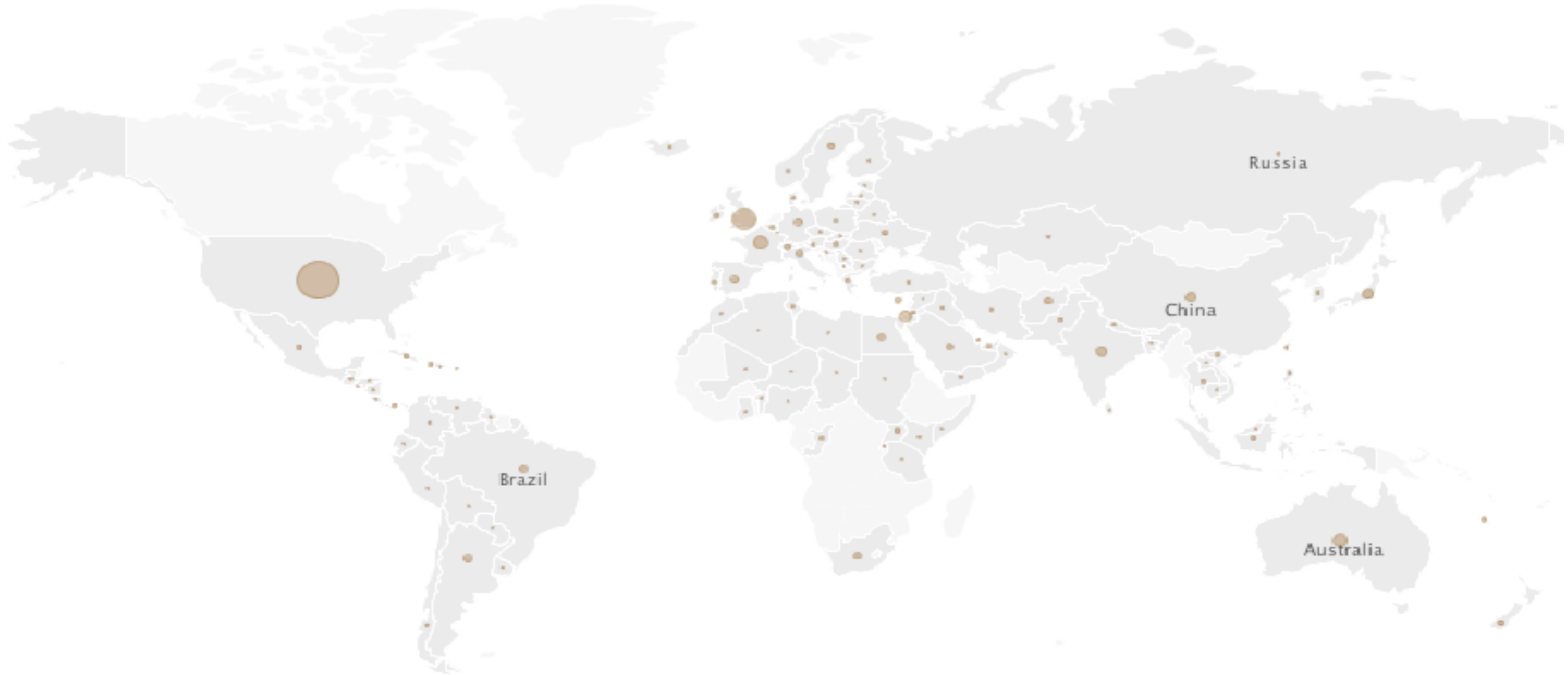
Robert Granke, executive director of CLWR, visited congregations of the ILSN in November. You can read more about his visit at www.lcccontheroad.ca, The Canadian Lutheran or in the forthcoming issue of CLWR's Partnership newsletter due out in early February.



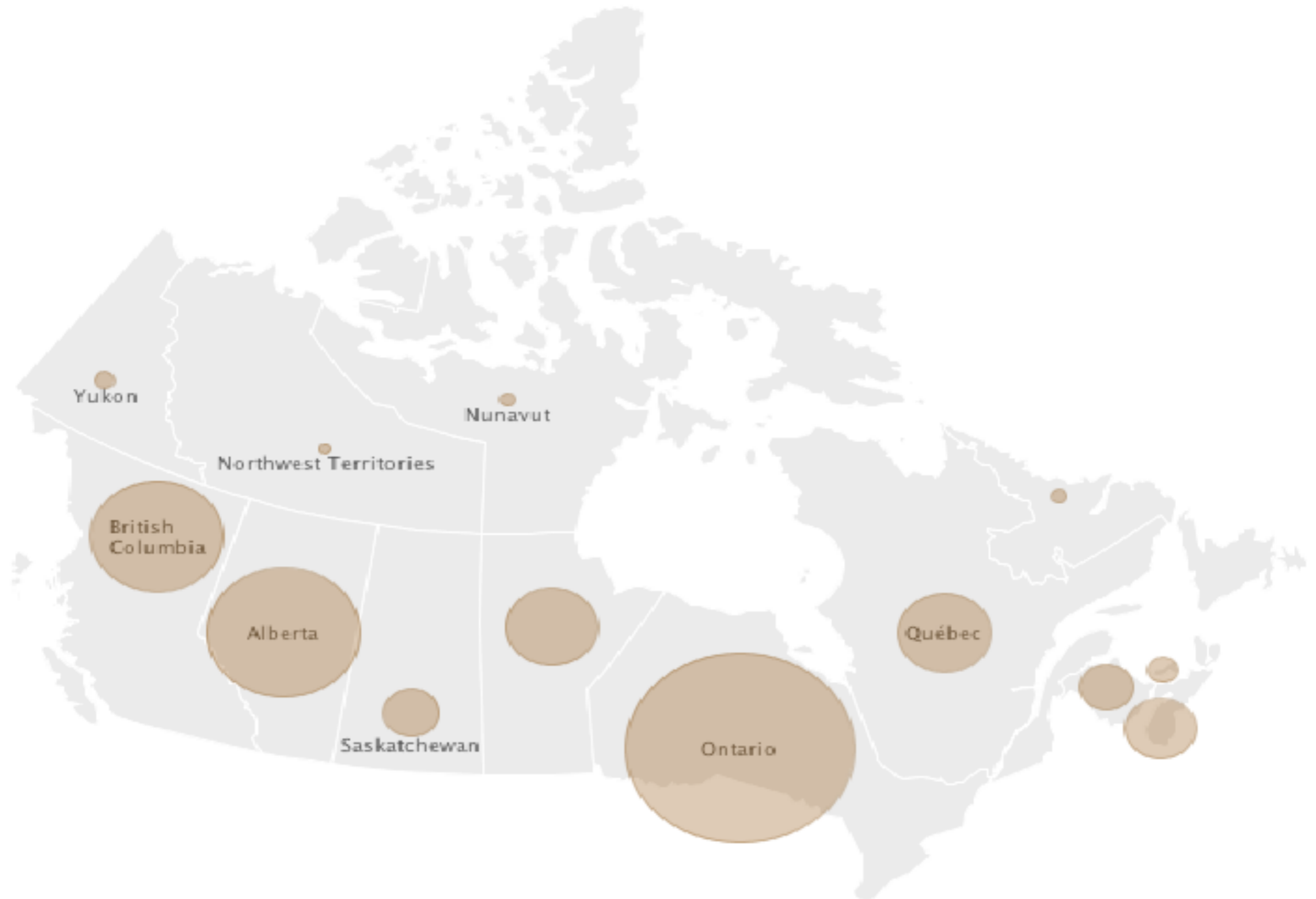
A medical clinic in Nicaragua.

Named Entity Recognition
as another approach?

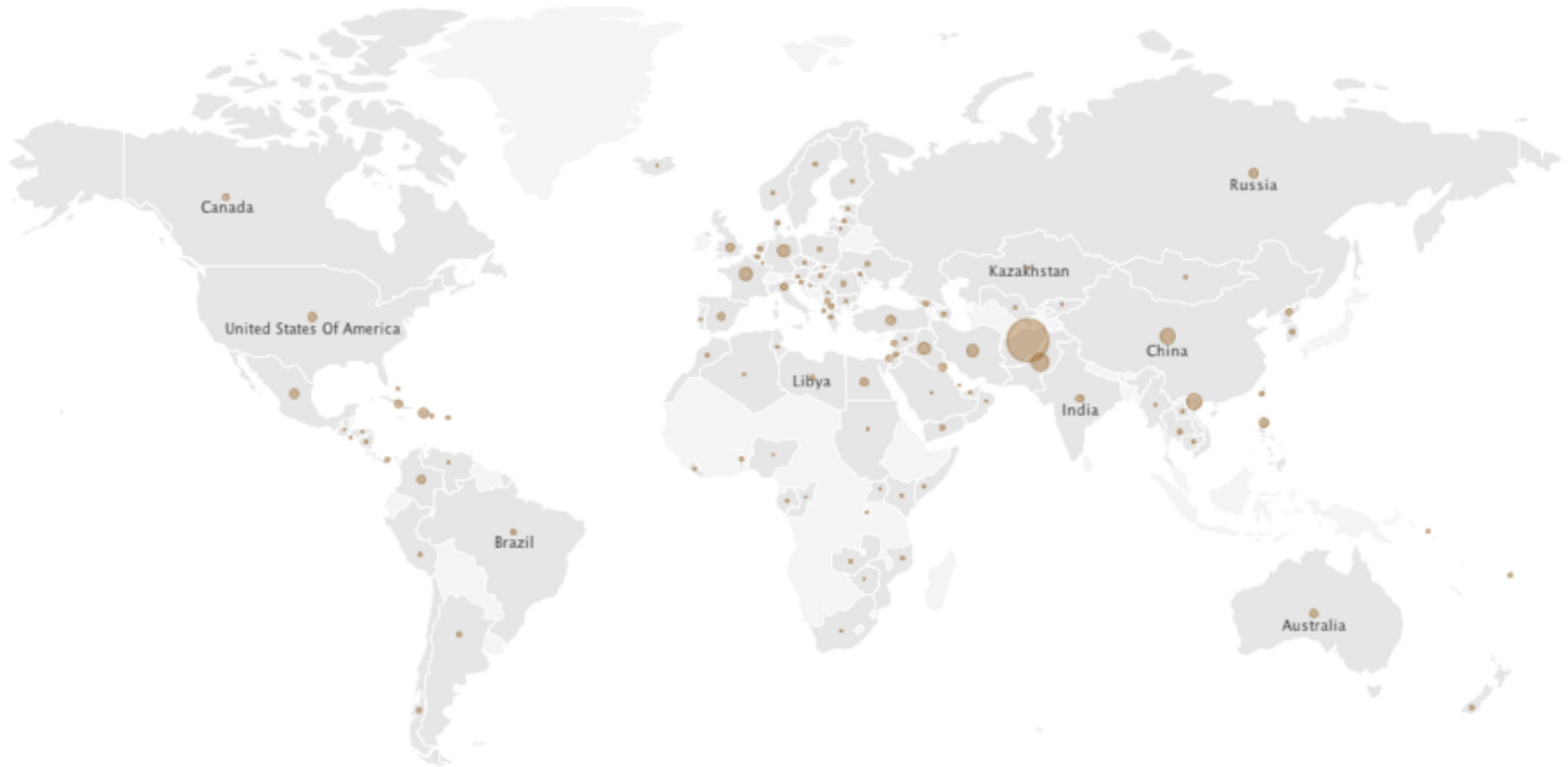
Countries Mentioned in .ca TLD (excluding Canada)



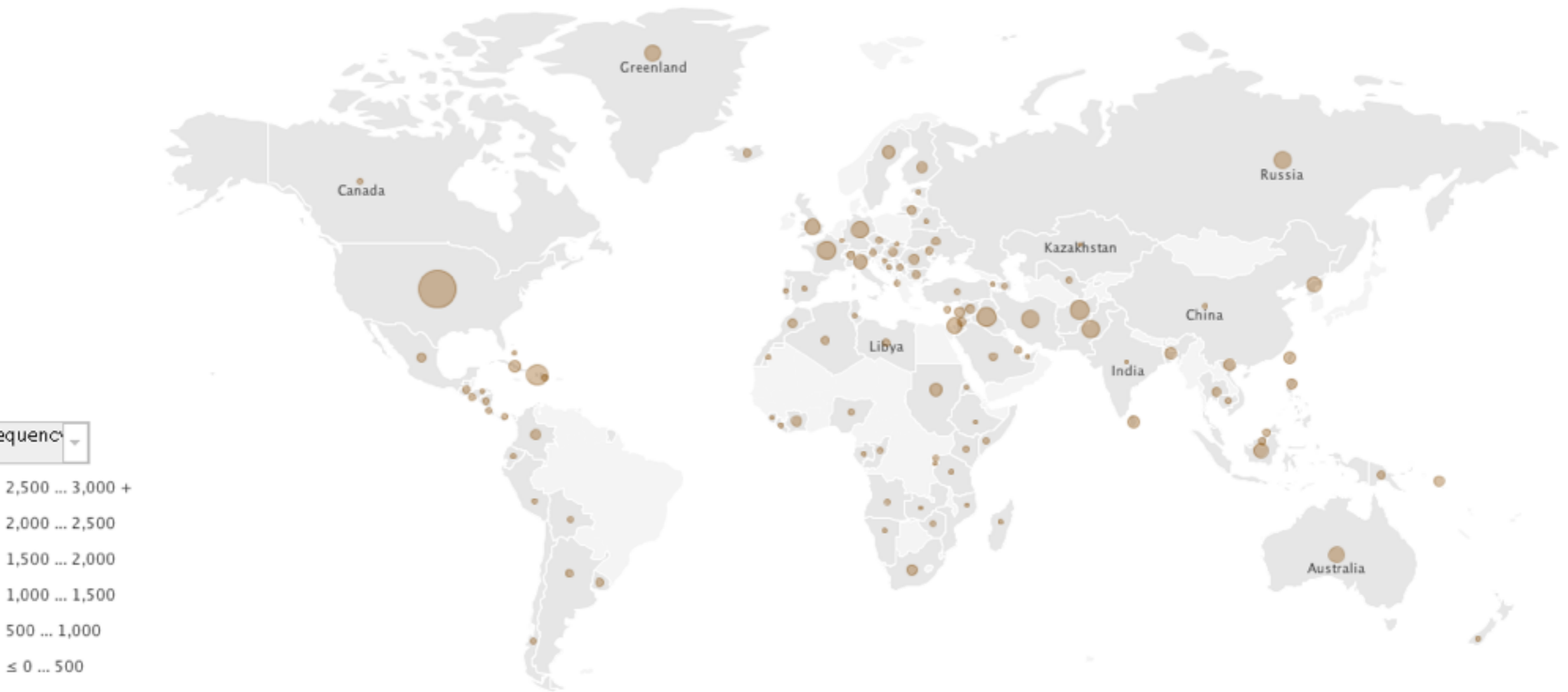
Provinces Mentioned in .ca TLD



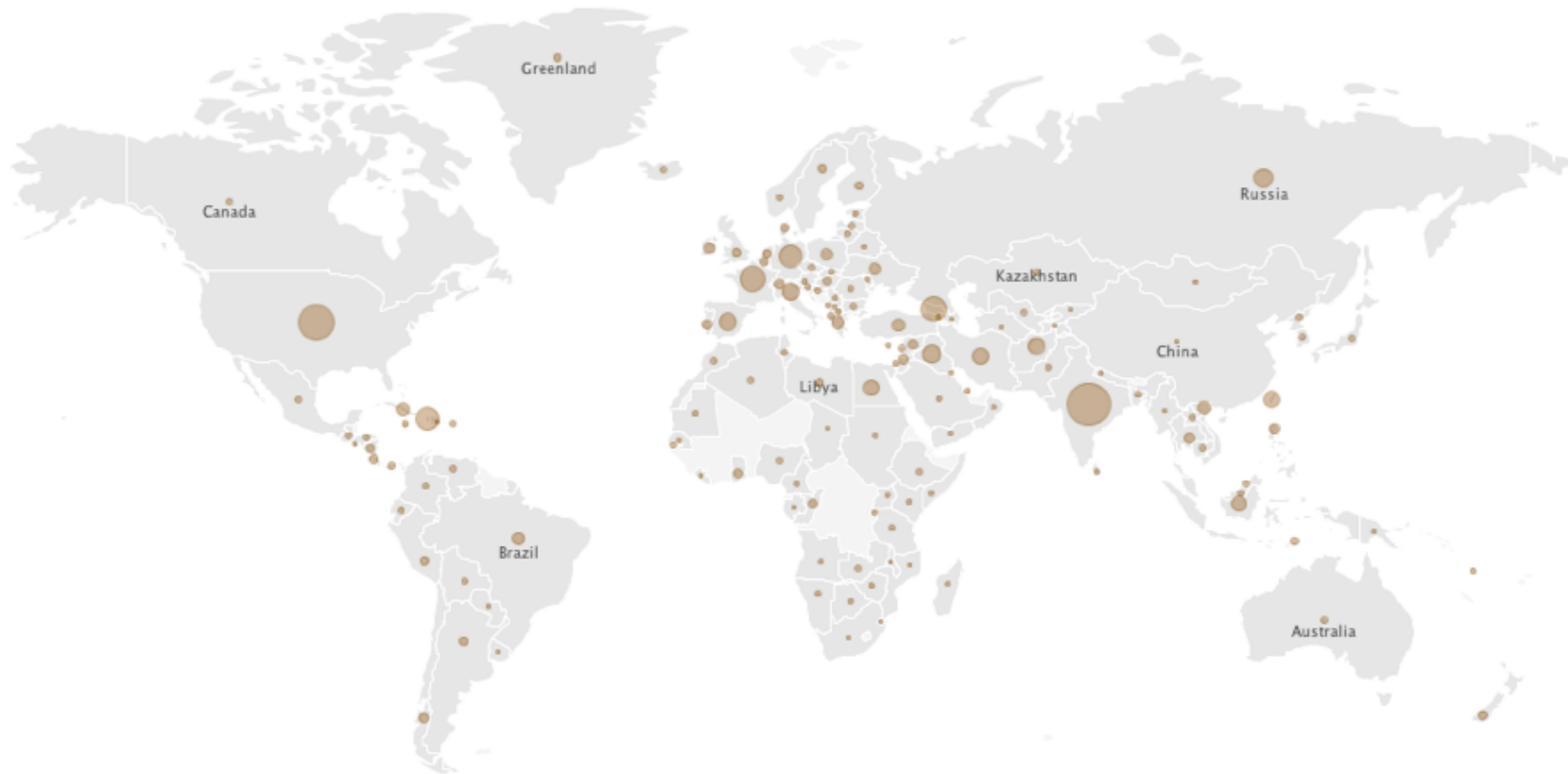
Countries Mentioned in .mil TLD



Countries Mentioned in .gov TLD



Countries Mentioned in .edu TLD



Frequency

10,000 ... 12 K +

8,000 ... 10,000

6,000 ... 8,000

4,000 ... 6,000

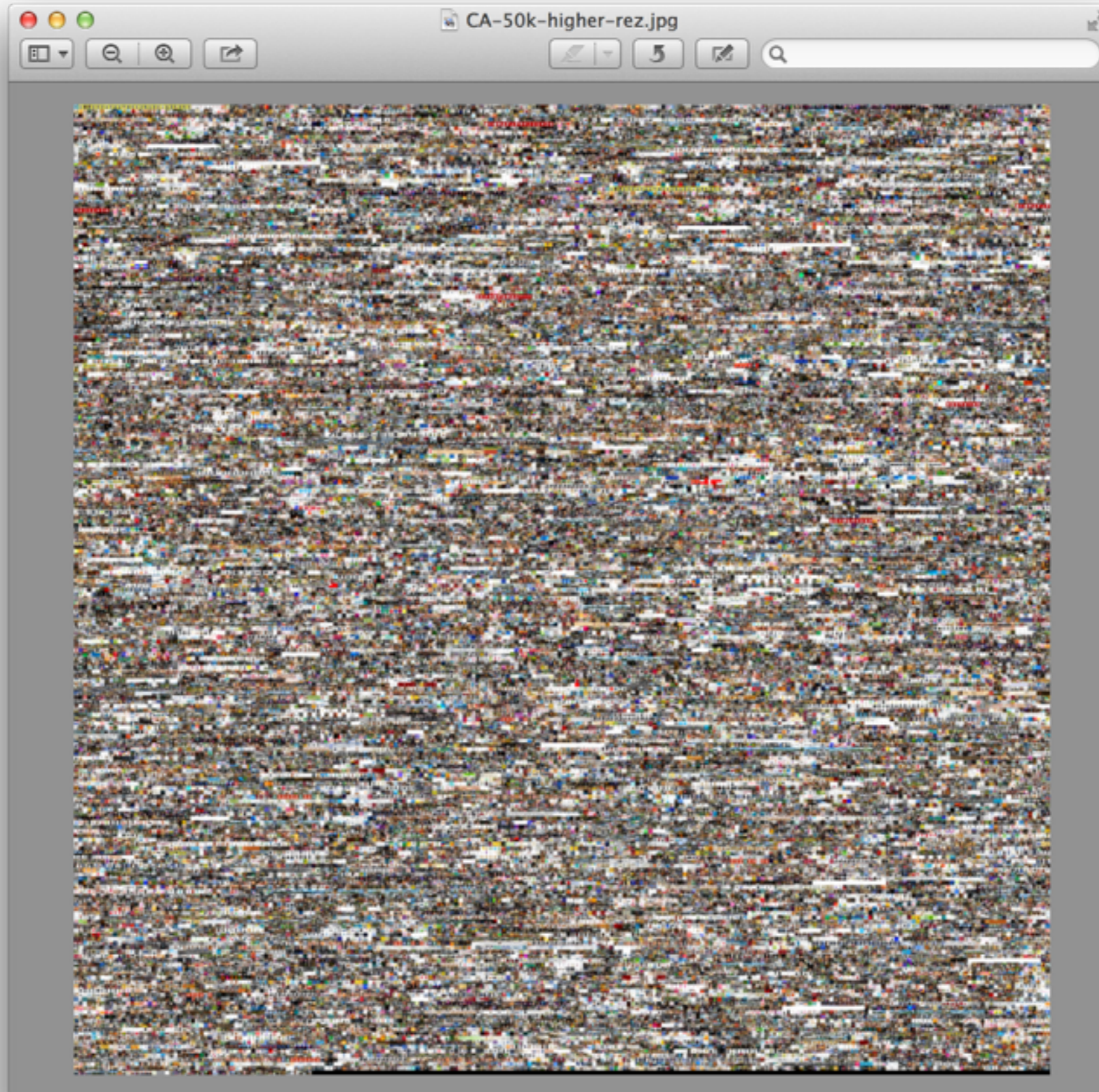
2,000 ... 4,000

≤ 0 ... 2,000

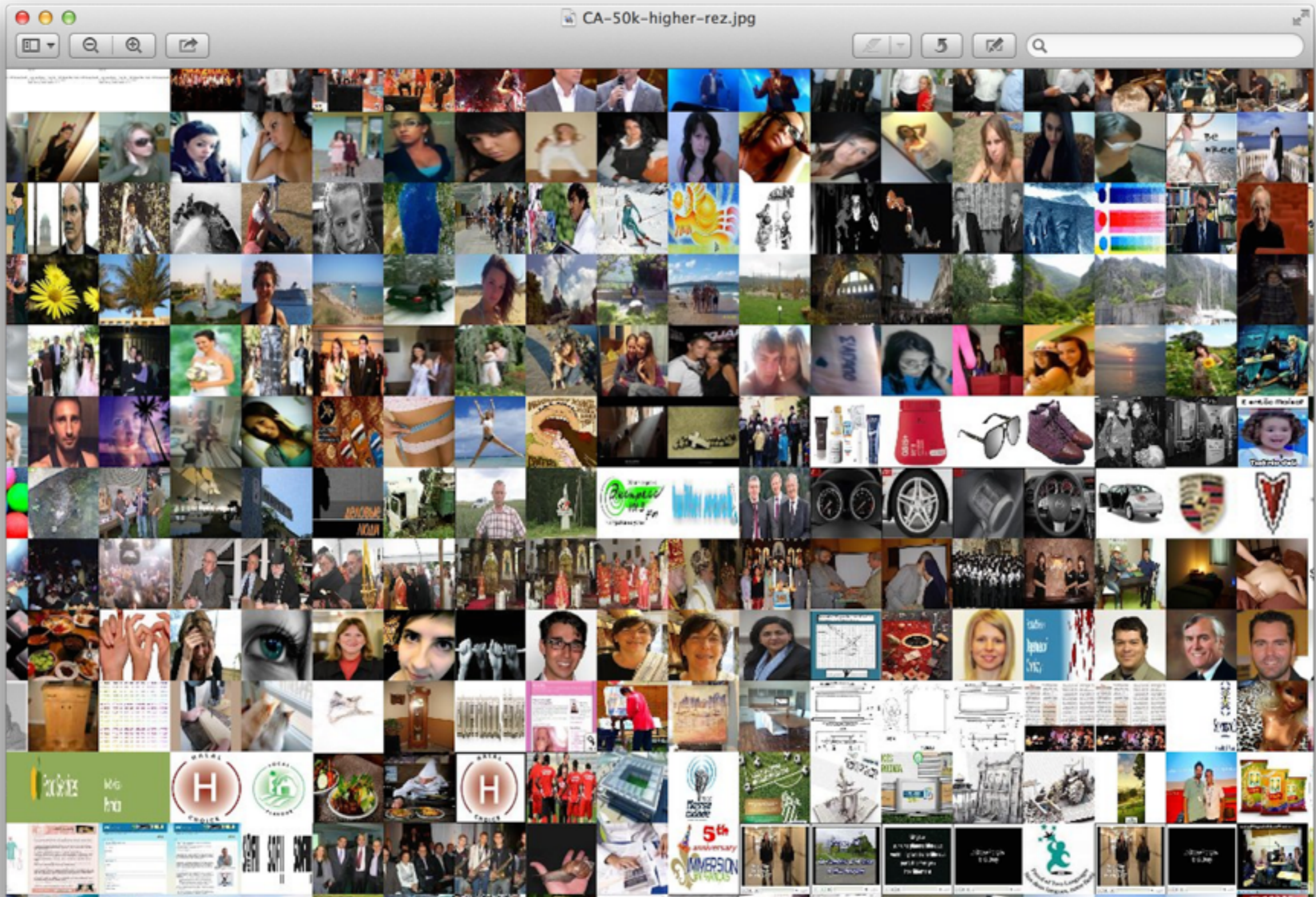
Countries Mentioned in .uk TLD (excluding UK)



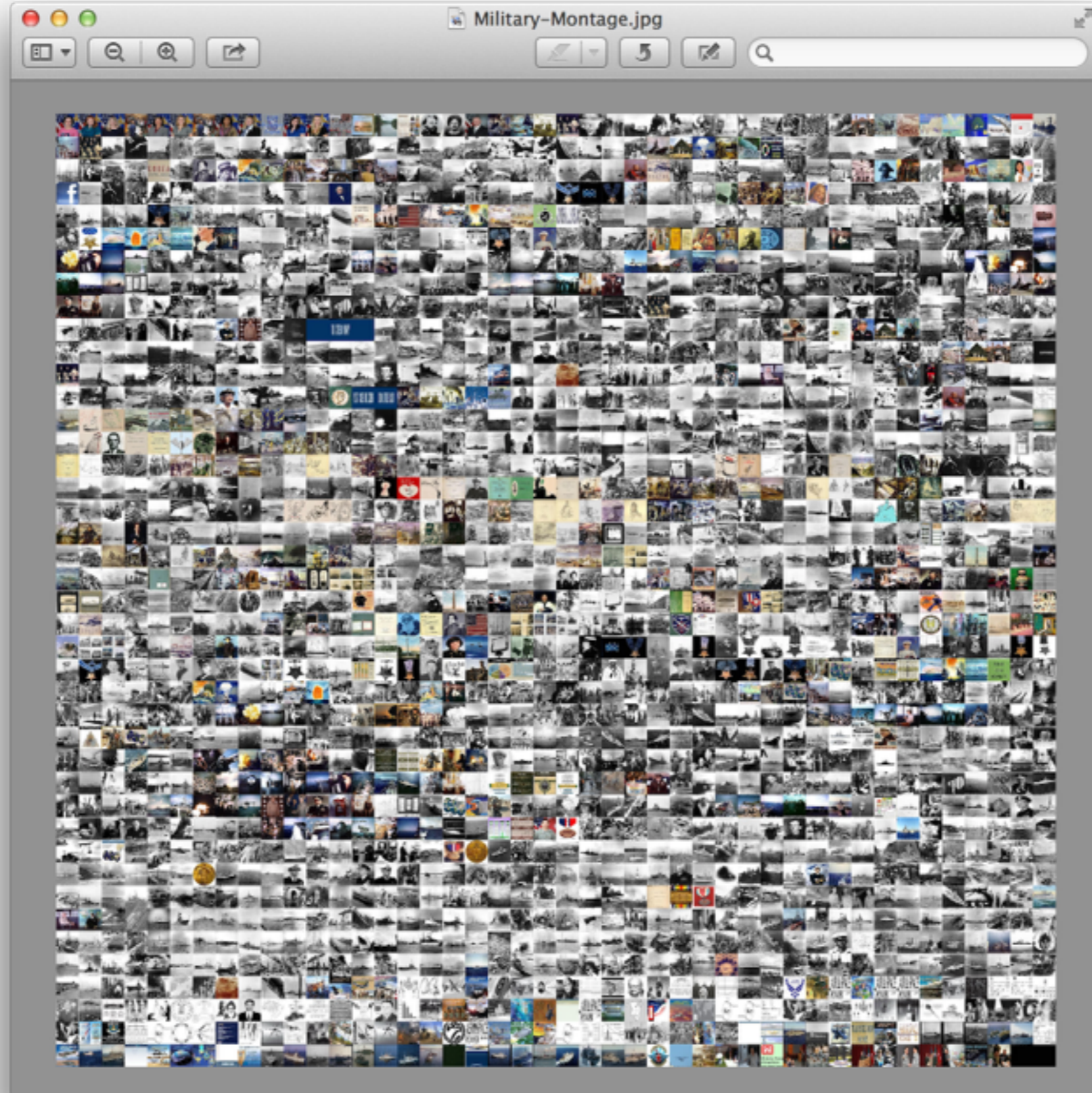
.ca montage



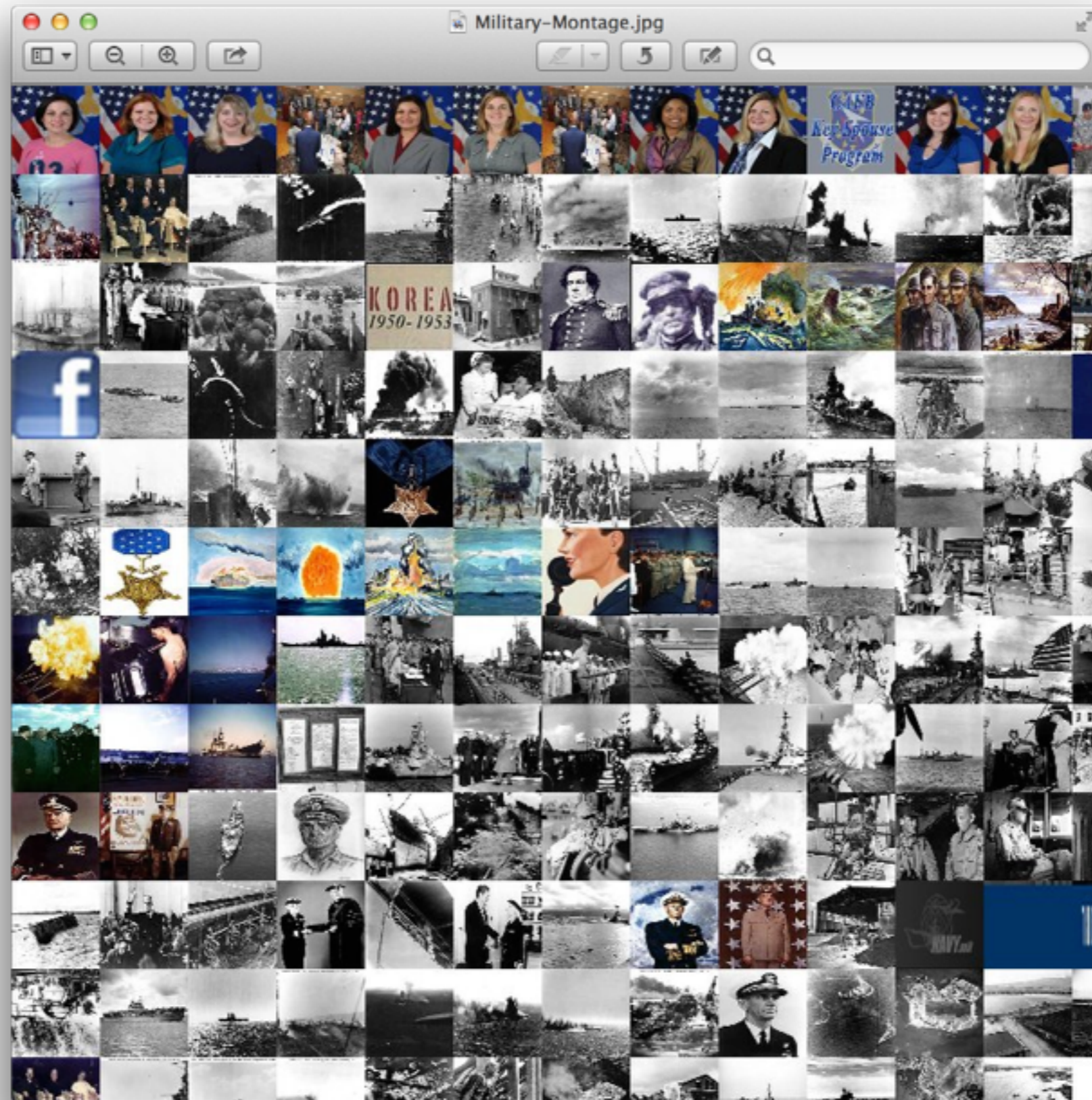
.ca montage (zoomed in)



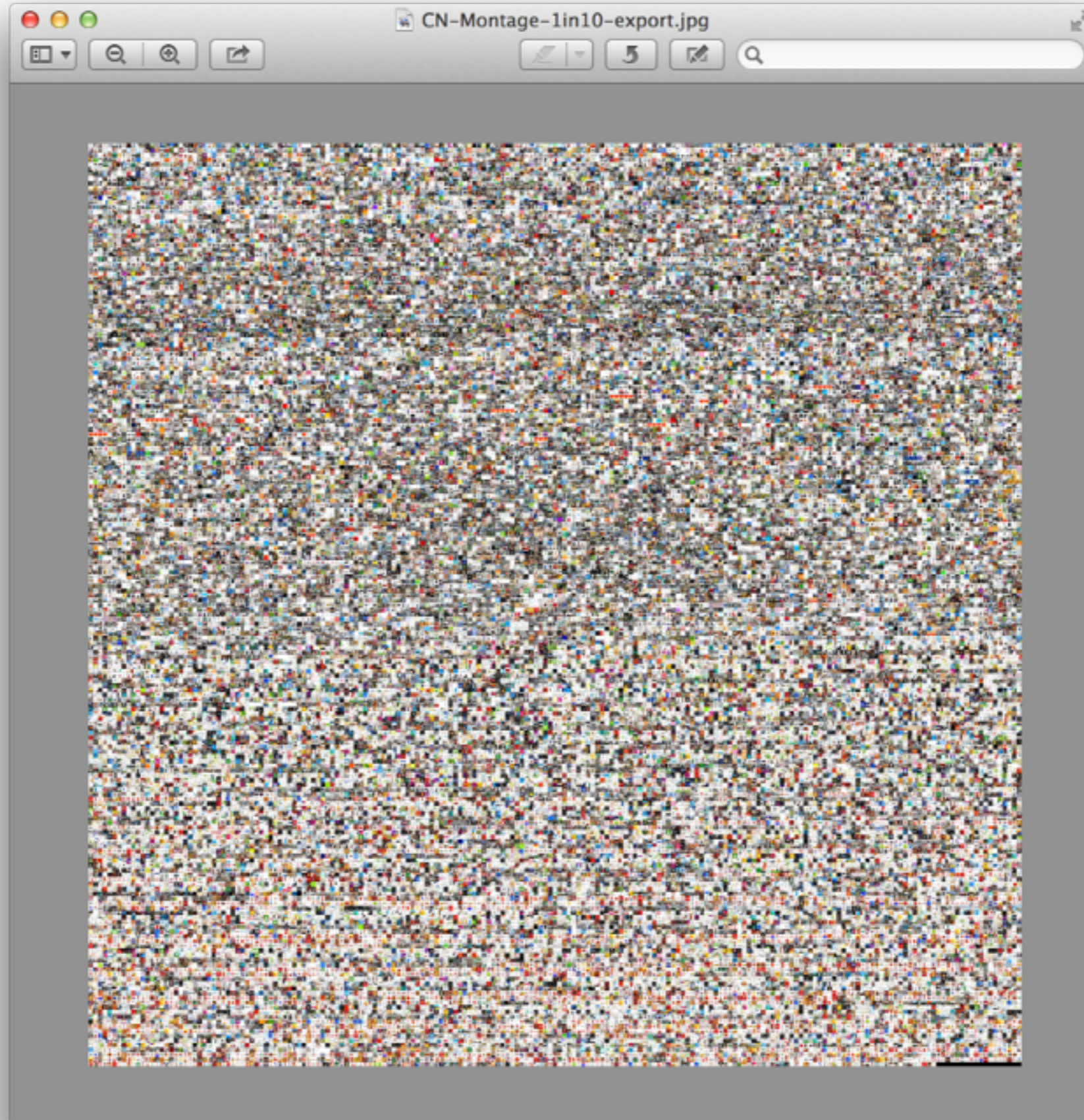
.mil montage



.mil montage (zoomed in)



.cn montage



.cn montage (zoomed in)



Thank you!

i2milligan@uwaterloo.ca

Ian Milligan, PhD
Assistant Professor of History
i2millig@uwaterloo.ca



UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada