

How big can a static site be?

Staticizing a census database

Martin Holmes and Greg Newton
University of Victoria Humanities Computing and Media Centre



University
of Victoria

Humanities Computing
and Media Centre



Project Endings and static websites: Why?

According to our research the long-term viability of DH projects is tenuous at best

- More than half of all DH projects do not have long-term preservation plans
- More than half of all DH projects do not include planning for an endpoint
- Only 10% of all DH projects consider their documentation to be adequate



The long-term viability of DH projects is tenuous at best



- Over 20% of DH projects stop working due to software obsolescence
- Responsibility for long-term maintenance falls to PI or nobody in nearly half of all DH projects
- Ongoing funding for maintenance is a major obstacle



Project Endings and static websites: Why?

Restoring/updating old projects is time-consuming and costly.

Project Endings offers practical solutions for obviating the need for long-term maintenance.

Is your project viable in the long term?

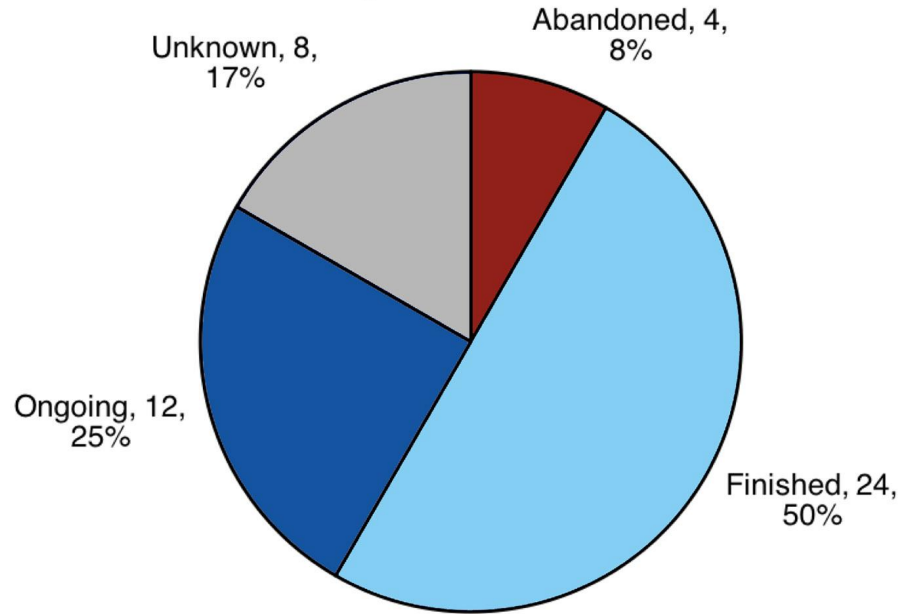
<https://hcmc.uvic.ca/endings/questionnaire.htm>



University
of Victoria

Humanities Computing
and Media Centre

Project status



2019 status of 48 DH2005 projects that had a web component

[The Final Death\(s\) of Digital Scholarship. Davis \(2019\)](#)



University
of Victoria

Humanities Computing
and Media Centre

Project has the most which is most content with the least.

– Diogenes

Be like Diogenes and throw away the cup

- Keep: HTML, CSS, Javascript
- Avoid: external dependencies
- Discard databases
- Say no to CMSs



University
of Victoria

Humanities Computing
and Media Centre

Project Endings and static websites: How?

Perceived realities

- Applications needs maps, graphs and all sorts. Can I avoid third-party tools?
- I need a database! / Where does my data live?
- How do I search my data?
- Can big applications operate with such constraints?



Project Endings and static websites: Working examples

The Map of Early Modern London (13,086 pages: mapoflondon.uvic.ca)

The Colonial Despatches (10,826 pages: bcgenesis.uvic.ca)

Digital Victorian Periodical Poetry (20,685 pages: dvpp.uvic.ca)



Our main focus has been *doing this for real to demonstrate that it is practical.*

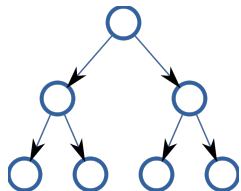


University
of Victoria

Humanities Computing
and Media Centre

But is there a limit?

Document collections such as digital editions need:



browsability (drill down through a hierarchy to find documents)



searchability (look for text, filter results based on dates, document types and so on)

Is there a limit on the scale of project which can be staticized, and if so, what is it?

We searched for a candidate to push the boundaries.



VIHistory (Vancouver Island History)

- PostgreSQL/PHP project
- 15 years old
- Primarily census data from Vancouver Island
- Census records from 1871, 1881, 1891, 1892*, 1901, 1911
- Around 150,000 records
- Associated tables of occupations, familial relationships, locations, addresses, religions, languages, nationalities and more.
- Already partially broken due to PHP and DB updates.



Challenges:

- The quality of the data
- The nature of census data itself
- The organization of the data in the db



Quality of the data

RECEMENT 1881 - CENSUS

RECEMENT 1881 - CENSUS

Province of *British Columbia*

District No. *190*

S District *Johnson St. West*

SCHEDULE No. 1 - Nominal Return of the Living.

TABLEAU No. 1 - Énumération des Vivants.

PAGE *92*

NOMINAL RETURN OF THE LIVING						NOMINAL RETURN OF THE LIVING										NOMINAL RETURN OF THE LIVING									
NAME	AGE	SEX	DATE OF BIRTH	PLACE OF BIRTH	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE	RELIGION	EDUCATION	PROFESSION, OCCUPATION OR TRADE
<i>Calhoun Ann</i>	<i>7</i>	<i>2</i>	<i>1874</i>	<i>BC</i>	<i>Presbyterian</i>	<i>Irish</i>																			
<i>Calhoun Horatio</i>	<i>33</i>	<i>✓</i>	<i>1841</i>	<i>BC</i>	<i>Catholic</i>	<i>German</i>	<i>Miner</i>	<i>M</i>																	
<i>Calhoun Ann</i>	<i>29</i>	<i>✓</i>	<i>1852</i>	<i>BC</i>		<i>Irish</i>																			
<i>Calhoun John</i>	<i>21</i>	<i>✓</i>	<i>1860</i>																						
<i>Calhoun John</i>	<i>19</i>	<i>✓</i>	<i>1862</i>																						
<i>Calhoun William</i>	<i>17</i>	<i>✓</i>	<i>1864</i>																						
<i>Calhoun Marie</i>	<i>15</i>	<i>✓</i>	<i>1866</i>																						
<i>Calhoun Robert</i>	<i>1</i>	<i>✓</i>	<i>1874</i>	<i>BC</i>																					
<i>Calhoun Jane</i>	<i>33</i>	<i>✓</i>	<i>1841</i>	<i>BC</i>	<i>Presbyterian</i>	<i>Irish</i>	<i>Saloon keeper</i>	<i>M</i>																	
<i>Calhoun Marie</i>	<i>21</i>	<i>✓</i>	<i>1853</i>			<i>American</i>																			
<i>Calhoun Margaret</i>	<i>4</i>	<i>✓</i>	<i>1874</i>	<i>BC</i>		<i>Irish</i>																			
<i>Calhoun Marie</i>	<i>2</i>	<i>✓</i>	<i>1872</i>																						
<i>Calhoun Anne</i>	<i>72</i>	<i>✓</i>	<i>1802</i>	<i>Scotland</i>	<i>Presbyterian</i>	<i>Irish</i>		<i>M</i>																	
<i>Calhoun George</i>	<i>58</i>	<i>✓</i>	<i>1816</i>	<i>Scotland</i>	<i>Presbyterian</i>	<i>Irish</i>		<i>M</i>																	
<i>Calhoun Joseph</i>	<i>52</i>	<i>✓</i>	<i>1820</i>					<i>M</i>																	
<i>Calhoun Thomas</i>	<i>20</i>	<i>✓</i>	<i>1854</i>																						
<i>Calhoun David</i>	<i>15</i>	<i>✓</i>	<i>1859</i>	<i>BC</i>																					
<i>Calhoun Frank</i>	<i>26</i>	<i>✓</i>	<i>1848</i>	<i>BC</i>	<i>Presbyterian</i>	<i>Irish</i>	<i>Transporter</i>	<i>M</i>																	
<i>Calhoun Mary</i>	<i>24</i>	<i>✓</i>	<i>1854</i>			<i>German</i>																			
<i>Calhoun Harold</i>	<i>12</i>	<i>✓</i>	<i>1872</i>	<i>BC</i>		<i>Irish</i>		<i>1</i>																	
<i>Calhoun William</i>	<i>26</i>	<i>✓</i>	<i>1846</i>	<i>USA</i>	<i>Ch. Engle</i>	<i>American</i>	<i>Calmet maker</i>	<i>M</i>																	
<i>Calhoun Frances</i>	<i>24</i>	<i>✓</i>	<i>1854</i>			<i>English</i>																			
<i>Calhoun Ellen</i>	<i>4</i>	<i>✓</i>	<i>1874</i>	<i>BC</i>		<i>American</i>																			



Province of Johnson District No. 190
Eager Lawrence

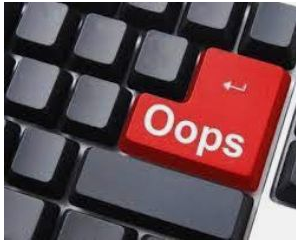
SCHEDULE No. 1—Nominal Return of the Living.
 TABLEAU No. 1—Énumération des Vivants.

8 District Johnson RE-CENSSEMENT 1881—CENSUS.
 PAGE 92

Description of the Names of the Persons						NAME	SEX	AGE	Date within last year	Country or Province of Birth	RELIGION	ORDRE	Profession, Occupation or Trade	Marital or Widowed	Instruction				Date of Emigration and Remarks
French	English	Native	Native	Native	Native										Reading	Writing	Arithmetic	Other	
Description of the Names of the Persons						NAME	SEX	AGE	No. days in last year	Days in Province of Birth	RELIGION	ORDRE	Profession, Occupation or Trade	Marital or Widowed	Instruction				Date of Emigration and Remarks
French	English	Native	Native	Native	Native										Reading	Writing	Arithmetic	Other	
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
						667/668	Carlson Ann	F	25	✓	N.B.	Presbyterian	Irish						May 12. 1871. 1st child
							Carlson Horatio	M	53	✓	N.B.	Catholic	German	G. Minis	M				
							Moore Ann	F	50	✓	N.B.		Irish						
							Moore George	M	29	✓			German						1st child. 1871.
							Edwards H.	M	21	✓			Shoreman						
							Horner	F	19	✓			Irish male						
							McLennan	M	17	✓									
							McMullen	F	15	✓			Irish male						
							McMullen	M	6	✓									
							Robert	M	1	✓	N.B.								
						677/685	White James Al.	M	33	✓	Scotland	Presbyterian	Scottish	Seton Kirk	M				
							White Minnie	F	21	✓	USA		American						
							Harvard	M	4	✓	N.C.		Scottish						
							Minnie	F	2	✓									
							Henry Al.	M	2	✓									
						678/686	Pottinger George	M	56	✓	Scotland	Presbyterian	Scottish		M				
							Isabella	F	52	✓									
							Thomas	M	20	✓									
							David	M	15	✓	N.B.								
						679/687	Burns Frank H.	M	26	✓	N.S.	Presbyterian	Irish	Landryman	M				
							May	F	24	✓	USA		German						
							Harold H.	M	12	✓	N.C.		Scottish						
						680/688	Forbes William	M	24	✓	USA	Ch. Eng.	American	Cabinet maker	M				
							Frederic	F	25	✓	USA		English						
							Ellen E.	F	4	✓	N.C.		American						

Opportunities for error

- Original enumerators make mistakes
- Transcribers working from grainy microfiche make mistakes
- Entry into spreadsheets with no data-constraints adds more
- Ingestion from spreadsheets into database creates yet more



So:

Values for gender
(should be “M” or
“F”)

- 0
- 9
- S
- D
- !

Values for hourly
wage (1911):

- \$40.00
- \$100.00
- \$500.00
- \$3,000.00



Challenges of census data

Each census collects information different from the last, because the preoccupations of society and government change.



The 1871 census is obsessed with race:

There are fields for counts of how many of these are in the household:

- white male
- white female
- chinese male
- chinese female
- colored [sic] male
- colored female
- native male
- native female



Chinese houseboys for the Kenneth McKenzie family at Craigflower and Lakehill



In 1891, there's a sudden interest in living conditions:

- building type (shanty, house, hotel)
- construction (wood, brick, stone, longhouse)
- number of floors
- number of rooms



506 Government Street, Hon. John Robson, M.P., residence, 1891



By 1911, the focus is on work and money:



City of Victoria Archives

Woodworkers Ltd. Sash and Door Factory, 2843 Douglas Street, c. 1912

- occupation
- earnings
- employment, employment state, other employment
- weeks working
- hours working
- weekly wage, hourly wage
- insurance (life and health)



University
of Victoria

Humanities Computing
and Media Centre

This variability was reflected in the DB structure...

Every census was in a separate table.

Every census table had different fields.

Linked tables for common values (location, nationality, religion) were chaotic because these values were expressed differently from year to year.

Most fields were “text” because the incoming data was so variable.



...and in the web application interface

Every census had a separate search page with a different interface.

There was a global search that attempted to schmush everything together, but omitted one entire census by accident, or perhaps in despair.

Searching was only possible within specific fields, not across an entire record.



Stage 1: convert to XML

We wrote a custom XML schema.

We have datatypes such as “moneyAmountOrGarbage”:

```
<dataSpec ident="vih.moneyAmountOrGarbage" module="vihistory">
  <desc>An amount of money in dollars and optional cents, or some garbage
from the source.</desc>
  <content>
    <dataRef name="token" restriction="(\d+(\.\d\d)?)|(QUERY: .+)" />
  </content>
</dataSpec>
```



One content model to rule them all

All entries use the same content model, with optional components and flexible structures smoothing out the differences between census years:



```

<entry date="1881" legacyId="56137" xml:id="cr_1881_56137">
  <title>Arme, (30), 1881, Victoria (190), Victoria City Johnson Street Ward (B)</title>
  <person gender="Female" ageYears="30" attendingSchool="false">
    <persName>
      <familyName>Arme</familyName>, <givenName/>
    </persName>
    <maritalStatus legacyId="1">Single</maritalStatus>
    <family ref="fam:f_1881_47_147">British Columbia, Victoria, Victoria City Johnson Street
      Ward</family>
    <religion legacyId="980">
      <desc genToken="rel_noneno_religion">None/No Religion</desc>
    </religion>
    <infirmities legacyId="0">Blank or None</infirmities>
    <event type="ownBirth">
      <place legacyId="15100">British Columbia, Canada (Native Indian/First Nations,
        Canada)</place>
    </event>
  </person>
  <location legacyId="47" srcType="fromCensus">
    <censusDistrict legacyId="190" date="1881">Victoria (190)</censusDistrict>
    <censusSubdistrict legacyId="B">Victoria City Johnson Street Ward (B)</censusSubdistrict>
  </location>
  <housing legacyId="2">
    <desc>House</desc>
  </housing>
  <work>
    <occupation legacyId="1057" code="X2100">
      <desc genToken="occ_none_or_unknown" srcType="fromCensus">None or Unknown</desc>
    </occupation>
  </work>
</entry>

```



Payoffs from data conversion

Removed over 500 completely empty records.

Removed about 4,500 mysterious unsourced records, possibly consisting of generated test data.

Normalized and corrected errors in hundreds of records.



Stage 2: Render to HTML

Granularity: 1 census record = 1 page

Browsability/drill-down: 1 data-point (a religion, a nationality, a family) = 1 listing page

Searchability: each census record has many <meta> tags from which **staticSearch** creates search filters (date, religion, nationality, location, race...)



Result

138,744 individual census record pages

166,528 pages across the site (because of generated listings pages)

Site size on disk: 3.6GB (of which 2.0GB is search index files)

Site build time: 1 hour 46 minutes (on Jenkins CI server)

staticSearch works fine – feel free to try it at the temporary site location:

<https://hcmc.uvic.ca/project/vicensus/>



Conclusions

~~We don't need no stinking database.~~ Even projects that look like a “natural” fit for a database may work better as static sites.

150,000 pages is not a big number for a static site, especially if they're small.

Sophisticated data constraints are easier in XML, so bad data is easier to control.

HOWEVER...



Conclusions (2)

We may be approaching a limit with the current **staticSearch**, because the index file containing document titles is approaching 16MB. This file has to be downloaded to the client before any search results can be shown, so the first search may appear to take a few seconds longer than subsequent searches.

We can imagine workarounds for this, involving changing the granularity (i.e. aggregating records into larger pages) at the project level, or splitting out the title file at the staticSearch build level.



Acknowledgements & links

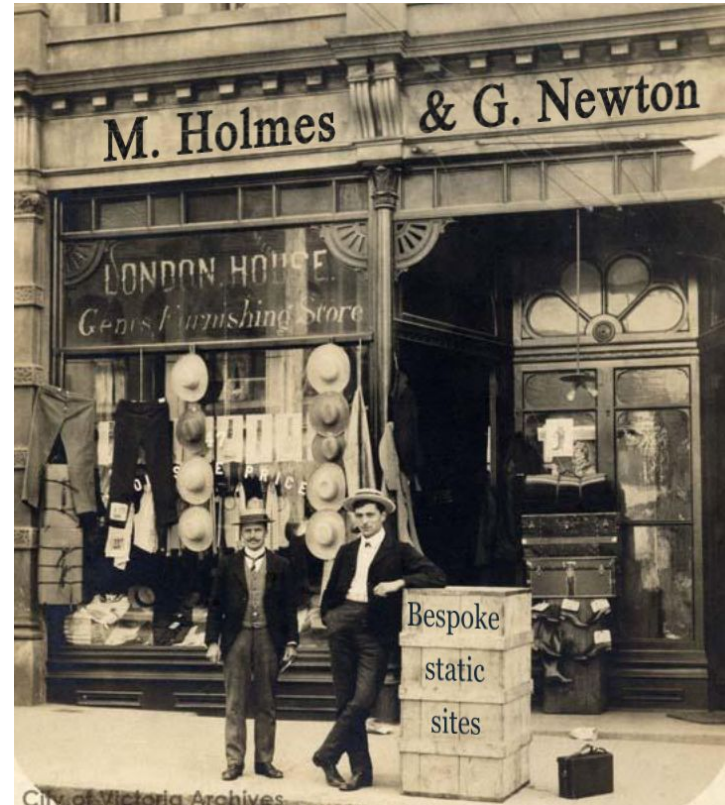
Thank you SSHRC for Endings funding.

Thank you Pat Dunae, VIHistory PI, for clear explanations and infinite patience.

Thank you Joey Takeda for co-authoring staticSearch.

endings.uvic.ca

github.com/projectEndings/staticSearch



University
of Victoria

Humanities Computing
and Media Centre