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

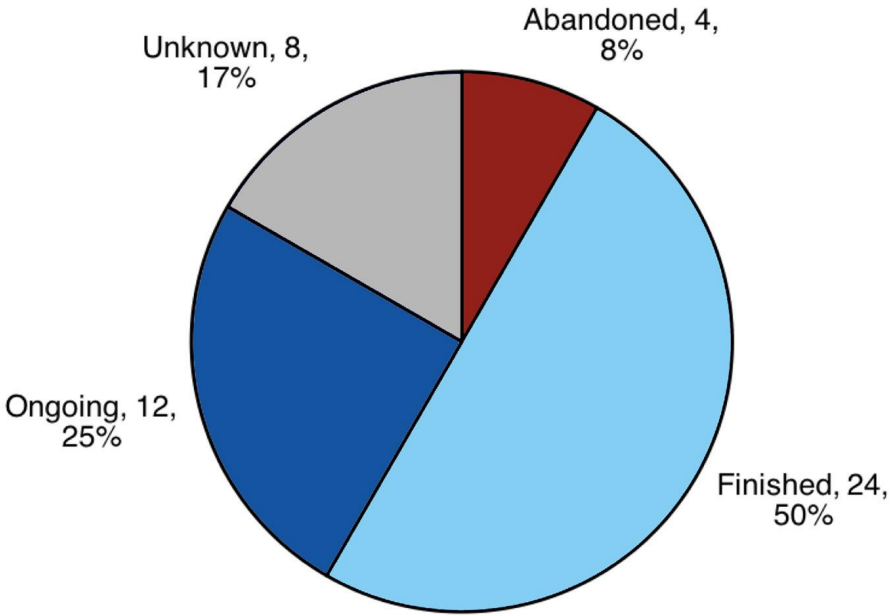
Project Endings and static websites: Why?

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

Project status



2019 status of 48 DH2005 projects that had a web component

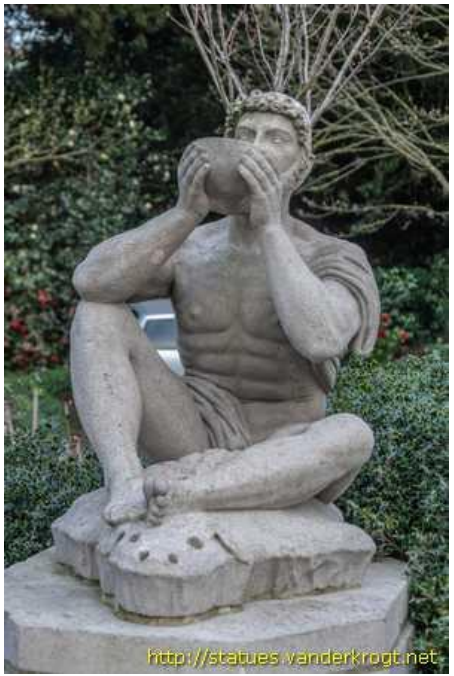
The Final Death(s) of Digital Scholarship, Davis (2019)

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



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)

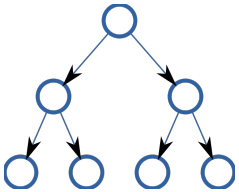
The Colonial Despatches (10,826 pages: bcgenesis.uvic.ca)

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.*

But is there a limit?

Document collections such as digital editions need:



browsability (drill down through a hierarchy to find documents)

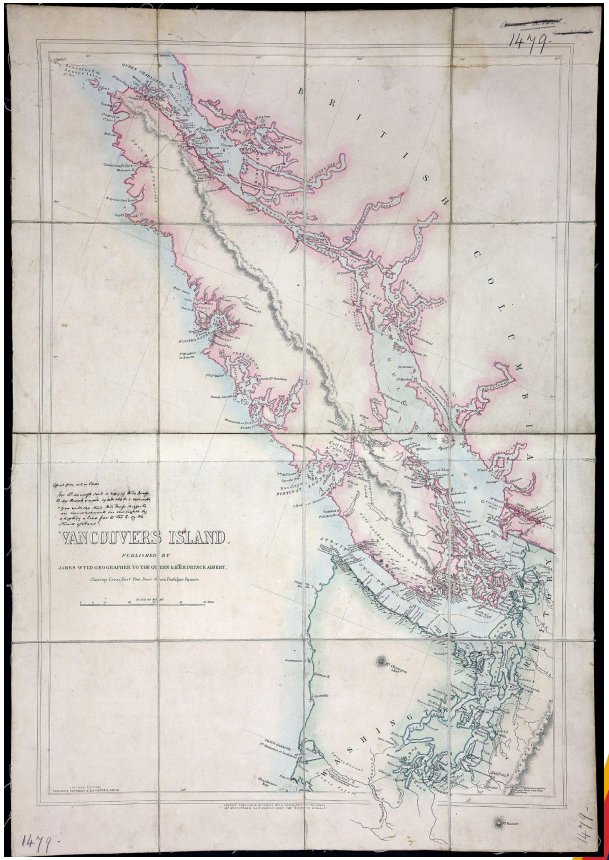
searchability (look for text, filter results based on dates, document types and so on)

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)

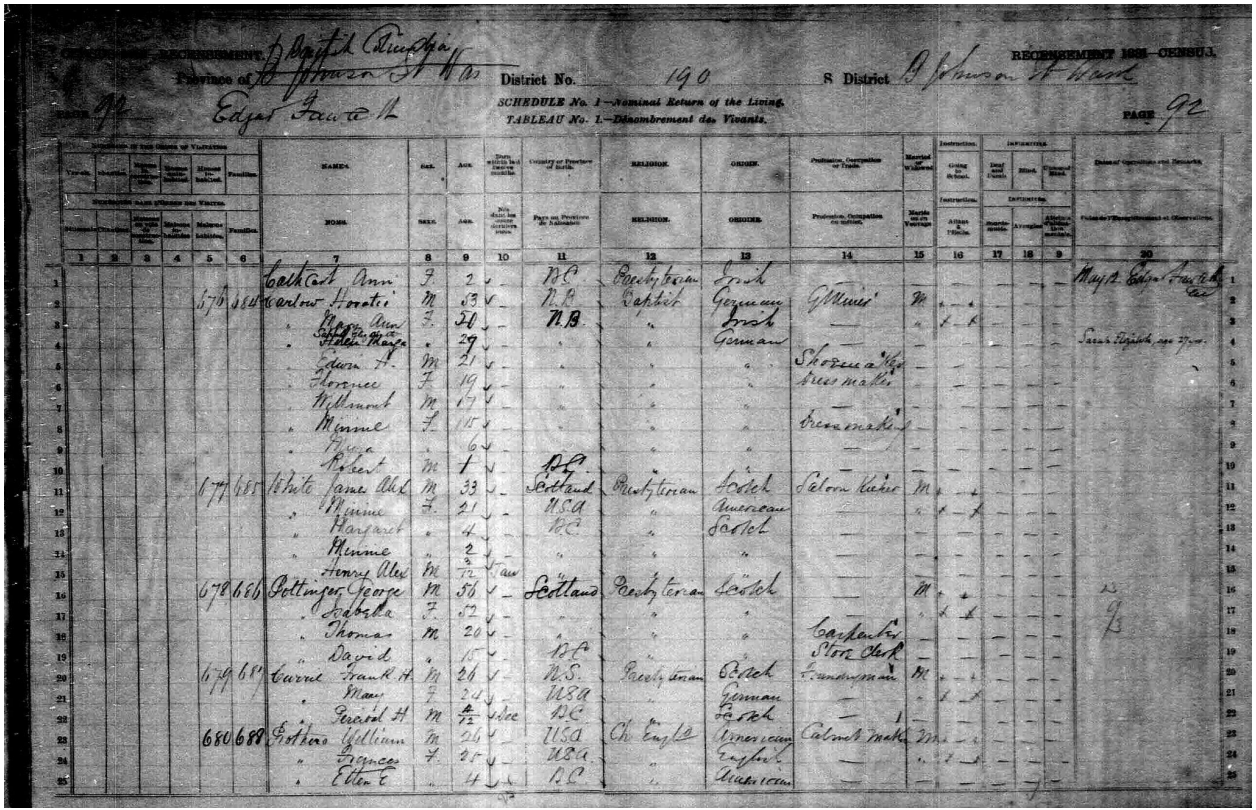
- PostgreSQL/PHP project
- 15 years old
- Primarily census data from Vancouver Island
- Census records from 1871, 1881, 1891, 1892*, 1901, 1911
- Around 150,000 records
- Associated tables of occupations, familial relationships, locations, addresses, religions, languages, nationalities and more.
- Already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



CENSUS—RECENSEMENT. Province of British Columbia — RECENSEMENT 1881—CENSUS.

D Johnson St Ward — District No. 190 — S District D Johnson St Ward

PAGE 92 — Edgar Fawcett — PAGE 92

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

	Vessels / Navires	Shanties / Chantiers	Houses in construction / Maisons en voie de construction	Houses uninhabited / Maisons inhabitées	Houses inhabited / Maisons habitées	Families / Familles	Names / Noms	Sex / Sexe	Age / Age	Born within last twelve months / Nés dans les douze derniers mois	Country or Province of Birth / Pays ou Province de Naissance	Religion / Religion	Origin / Origine	Profession, Occupation or Trade / Profession, Occupation ou métier	Married or Widowed / Mariés ou Veufs	Instruction: Going to School / Allant à l'École	Infirmities: Deaf and Dumb / Sourds-muets	Blind / Aveugles	Unsound Mind / Aliénés	Dates of Operations and Remarks / Dates de l'Enregistrement et Observations	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	9	20	
1							Cathcart Ann	F	2	—	B.C.	Presbyterian	Irish	—	—	—	—	—	—	May 12 Edgar Fawcett Enr	1
2					676	684	Carlow Horatio	M	53	—	N.B.	Baptist	German	G Miner	M	+ — +	—	—	—		2
3							" Mary Ann	F	50	—	N.B.	"	Irish	—	"	+ — +	—	—	—		3
4							" ~~Helen Maria~~ Sarah Elizabeth	"	27	—	"	"	German	—	—	—	—	—	—	Sarah Elizabeth, age 27 …	4
5							" Edwin A.	M	21	—	"	"	"	Shoemaker	—	—	—	—	—		5
6							" Florence	F	19	—	"	"	"	Dress maker	—	—	—	—	—		6
7							" Willmont	M	17	—	"	"	"	—	—	—	—	—	—		7
8							" Minnie	F	15	—	"	"	"	Dress making	—	—	—	—	—		8
9							" Mary	"	6	—	"	"	"	—	—	—	—	—	—		9
10							" Robert	M	1	—	B.C.	"	"	—	—	—	—	—	—		10
11					677	685	White James Alex	M	33	—	Scotland	Presbyterian	Scotch	Saloon Keeper	M	+ — +	—	—	—		11
12							" Minnie	F	21	—	U.S.A.	"	American	—	"	+ — +	—	—	—		12
13							" Margaret	"	4	—	B.C.	"	Scotch	—	—	—	—	—	—		13
14							" Minnie	"	2	—	"	"	"	—	—	—	—	—	—		14
15							" Henry Alex	M	3/12	Jan	"	"	"	—	—	—	—	—	—		15
16					678	686	Pottinger George	M	56	—	Scotland	Presbyterian	Scotch	—	M	+ +	—	—	—		16
17							" Isabella	F	52	—	"	"	"	—	"	+ +	—	—	—		17
18							" Thomas	M	20	—	"	"	"	Carpenter	—	—	—	—	—		18
19							" David	"	15	—	B.C.	"	"	Store Clerk	—	—	—	—	—		19
20					679	687	Currie Frank H.	M	26	—	N.S.	Presbyterian	Scotch	Draughtsman	M	+ — +	—	—	—		20
21							" Mary	F	24	—	U.S.A.	"	German	—	"	+ — +	—	—	—		21
22							" Gerald H	M	4/12	Dec	B.C.	"	Scotch	—	—	—	—	—	—		22
23					680	688	Gothers William	M	26	—	U.S.A.	Ch England	American	Cabinet Maker	M	— +	—	—	—		23
24							" Frances	F	25	—	U.S.A.	"	English	—	"	+ — +	—	—	—		24
25							" Ellen E	"	4	—	B.C.	"	American	—	—	—	—	—	—		25

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

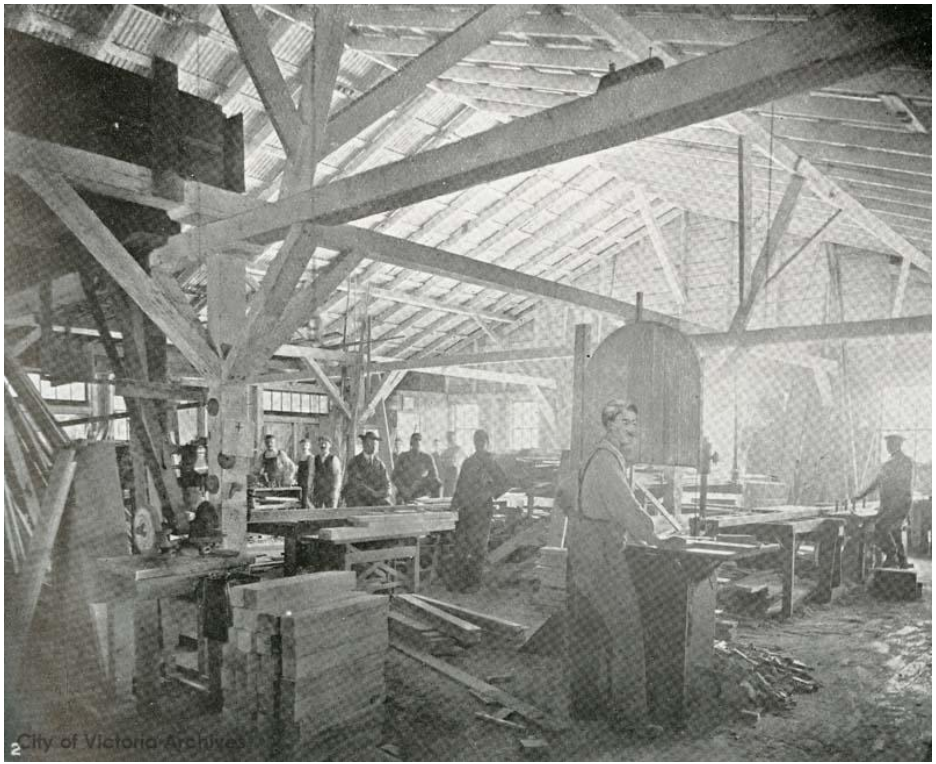
Chinese houseboys for the Kenneth McKenzie family at Craigflower and Lakehill

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.P., residence, 1891

By 1911, the focus is on work and money:



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)

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