When will I get a new logo?

Quarterly review Q1
October 2014

Parsoid Team: Core + extended

- Arlo, C.Scott, Marc, Subbu (all remote)
- Gabriel (now services)

With a little help from our friends

- Erica and other community liaisons
- VE, Flow, and other teams that use Parsoid
- Antoine, Ops, and many others @ WMF we (in)directly rely on …
- JackMcBarn, Bartosz (now with VE), Nemo, Nico V., and others who file bug reports, review code, help make sense of wikitext, …

Agenda

- Our objectives
- Progress Q1
- Tasks Q2
- Q & A

Our objectives

Photo by Nicolas Vigier - 2005-01-30, Creative Commons Attribution-Share Alike 2.0 Generic license
https://commons.wikimedia.org/wiki/File:Golden_gate.jpg

1. Faithful bidirectional conversion without dirty wikitext diffs
2. Improve performance & enable new features by moving our primary content representation to HTML5 + RDFa
3. Research platform for better templating, widget, and diffing solutions

Development context

- 2011-2012:
  - Is this bidirectional bridge feasible?
- 2012-2013:
  - Prototype to support VisualEditor
  - HTML5+RDFa spec for parsed output
- 2013-2014:
  - Production-ready and supports multiple clients
- 2014-2015:
  - Can Parsoid HTML be the content representation?
  - Can we enable other applications?

Background: Why is this hard?

Will skip this unless Damon/Lila/anyone else is interested in the details.

- Wikimania 2014 talk slides
- April 2014 Tech talk slides
- March 2013 blog post

Progress Q1

What we got done:

- Improved
  - rendering
  - roundtripping
  - Parsoid-specific CSS
  - robustness
  - testing infrastructure
- Visual diffing
- Wikitext linter GSoC project
- Started work on support for language variants
- OCG-based PDF rendering live

And what we didn't:

- Language variant editing support
- Support for certain types of templates
- Logging output to LogStash
- Stable id support yet to start
- Gallery support
- Researching content widgets

Progress Q1:

Continuous iteration

- Ongoing improvements to round-tripping & render accuracy
  - Reaaaaally loooong taaaiiiiiiiii
- Continuous deploys
  - mediawiki.org/wiki/Parsoid/Deployments
  - Very solid thanks to work invested in testing

Progress Q1:

Testing infrastructure: Parser tests

- Added HTML Tidy support
- Test results

	# tests	wt2html	wt2wt	html2wt	html2html	selser
Q4	1347	911 / 436	1198 / 125	390 / 895	814 / 467	16438 / 1198
Q1	1366	966 / 400	1222 / 115	437 / 855	909 / 392	16719 / 1137

- Ongoing work:
  - Separate true failures from testing framework limitations, incompatibilities, edge case WONTFIXes

Progress Q1:

Round-trip test results

- **100%** parsed without errors,
- **99.75%** round-tripped without semantic differences, and
- **85.15%** round-tripped with no character differences at all.

Q4: 0.25%

have semantic diffs without selser

We have run roundtrip-tests on **159440** articles, of which

Q1: 0.16%

have semantic diffs without selser

- **100%** parsed without errors,
- **99.84%** round-tripped without semantic differences, and
- **85.2%** round-tripped with no character differences at all.

0.16% = ~250 pages in RT-testing

0.16% = ~7000 pages on enwp

Latest revision:

Git SHA1	**9e793ebe02d0159783e64d7783e34a623518b10c**
Test Results	**158995**
Crashers	**1**
Fixes	**25**
Regressions	**25**
RT selser errors	**7**

Progress Q1:

Testing infrastructure: Visual diffs

- RT tests identify wt ↔ wt FIXMEs
  - Important for editing without dirty diffs

- Visual diffs identify rendering FIXMEs
  - Important for identifying semantic errors, page views from Parsoid HTML, WYSIWYG editing
  - Foundation for Parsoid-based PDF/Zim/ePub renders, a better “printable page”?

Progress Q1:

Testing infrastructure: Visual diffs

- Mass visual diff testing
  - http://parsoid-tests.wikimedia.org/visualdiff/
  - Extended + repurposed our RT-testing framework

We have run visual differences tests on **791** articles, of which

- **100%** tested without errors,
- **65.11%** showed less than 1% differences, and
- **4.3%** rendered with pixel-perfect accuracy.

Latest revision: **Deployed Parsoid version**

Git SHA1	**deed30b2448d4ad500c5d9c9ec6680ca0708da8f**
Test Results	**794**
Crashers	**0**
Fixes	**0**
Regressions	**0**

Progress Q1:

Testing infrastructure: Visual diffs

- Take PHP parser + Parsoid screenshots
- Compare screenshots & mark up diffs
- Based on PhantomJS + ResembleJS
- Examples:
  - Good: enwiki/Hampi
  - Bad: enwiki/IIHF_World_Women's_Championships
  - Minor: enwiki/Pizza_Connection_Trial

Progress Q1:
Parsoid HTML5 page views

parsoid-lb.eqiad.wikimedia.org/enwiki/San_Francisco?oldid=612886865

the tallest building in San Francisco

Climate

A popular quote incorrectly attributed to Mark Twain is "The coldest winter I ever spent was a summer in San Francisco".[77] [78] San Francisco's climate is characteristic of the cool-summer Mediterranean climate (Csb)[79] of California's coast, "generally characterized by moist mild winters and dry summers".[80] Since it is surrounded on three sides by water, San Francisco's weather is strongly influenced by the cool currents of the Pacific Ocean, which moderate temperature swings and produce a remarkably mild year-round climate with little seasonal temperature variation.

Fog is a regular feature of San Francisco summers.

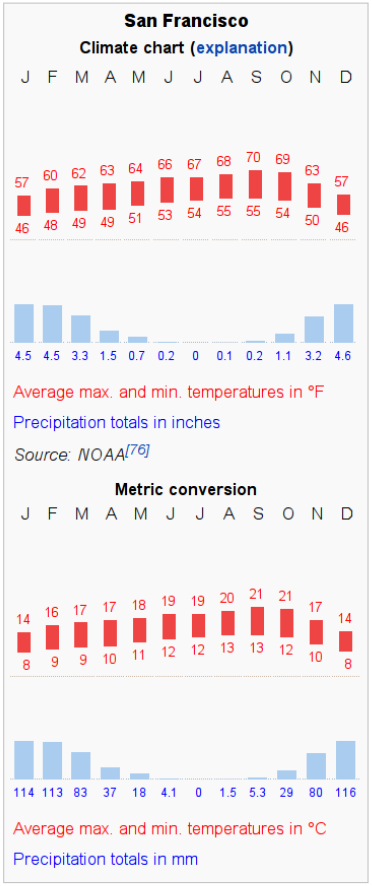
Among major U.S. cities, San Francisco has the coldest daily mean, maximum, and minimum temperatures for June, July, and August.[81] During the summer, rising hot air in California's interior valleys creates a low pressure area that draws winds from the North Pacific High through the Golden Gate, which creates the city's characteristic cool winds and fog.[82] The fog is less pronounced in eastern neighborhoods and during the late summer and early fall, which is the warmest time of the year.

Because of its sharp topography and maritime influences, San Francisco exhibits a multitude of distinct microclimates. The high hills in the geographic center of the city are responsible for a 20% variance in annual rainfall between different parts of the city. They also protect neighborhoods directly to their east ("banana belts" such as Noe Valley) from the foggy and sometimes very cold and windy conditions experienced in the Sunset District; for those who live on the eastern side of the city, San Francisco is sunnier, with an average of 260 clear days, and only 105 cloudy days per year.

Temperatures reach or exceed 80 °F (27 °C) on an average of only 21 and 23 days a year at downtown and SFO, respectively.[76] The dry period of May to October is mild to warm, with the normal monthly mean temperature peaking in September at 62.7 °F (17.1 °C).[76] The rainy period of November to April is slightly cooler, with the normal monthly mean temperature reaching its lowest in January at 51.3 °F (10.7 °C).[76] On average, there are 73 rainy days a year, and annual precipitation averages 23.65 inches (601 mm).[76] Variation in precipitation from year to year is high. In 2013, a record low 5.59 in (142 mm) of rainfall was recorded at downtown San Francisco, where records have been kept since 1849;[76] low records were shattered in much of the state. Snowfall in the city is very rare, with only 10 measurable accumulations recorded since 1852, most recently in 1976 when up to 5 inches (130 mm) fell on Twin Peaks.[83][84]

The highest recorded temperature at the official National Weather Service office was 103 °F (39 °C) on July 17, 1988, and June 14, 2000. The lowest recorded temperature was 27 °F (−3 °C) on December 11, 1932.[85] The National Weather Service provides a helpful visual aid[86] graphing the information in the table below to display visually by month the annual typical temperatures, the past year's temperatures, and record temperatures.

San Francisco falls under the USDA 10b Plant Hardiness zone.[87]



Climate data for San Francisco (downtown),[lower-alpha 1] 1981–2010 normals, extremes 1849–present

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °F (°C)	79 (26)	81 (27)	87 (31)	94 (34)	97 (36)	97 (36)	98 (37)	98 (37)	101 (38)	102 (39)	86 (30)	76 (24)	102 (39)
Average high °F (°C)	56.9 (13.8)	60.2 (15.7)	61.8 (16.6)	63.1 (17.3)	64.3 (17.9)	66.4 (19.1)	66.5 (19.2)	68.1 (20.1)	70.2 (21.2)	69.2 (20.7)	63.1 (17.3)	57.1 (13.9)	63.9 (17.7)
	45.7	47.5	48.5	49.2	51.0	52.8	54.1	55.1	55.1	53.7	50.1	46.1	50.7

Progress Q1:

Preparing for HTML5 page views

- Continued to tweak Parsoid-specific CSS
  - Found issues via visual diff testing
  - Cite customizations via CSS
  - Tweaked Vector CSS
  - Serving extension-specific CSS
  - More to be done

- Check out Jackmcbarn's gadget

  Add `importScript('User:Jackmcbarn/parsoidview.js');`

  to https://en.wikipedia.org/wiki/Special:MyPage/common.js

Progress Q1:

Preparing for HTML5 page views

- Remove inlined data-parsoid from HTML
  - New /v2/ API endpoints -- used by RESTbase
  - Reduces network traffic
- Q2 TODO:
  - Integrate with RESTBase
  - Move data-mw out of HTML

Progress Q1:

More robust

- Improved pathological parsing scenarios
  - On some pages, document parsed all the way to end of document before backtracking (because of “syntax errors”)
  - Fixed tokenizer backtracking cache
  - These pages now parse in reasonable time

- Continued to fix edge case crashers found in production logs

Progress Q1:

More robust

- Improved handling of large tables
  - Sep 17: Parsoid cluster seized up (all 24 servers)
  - Traced it to a huwiki page with a 3000+ row table
  - Parsoid was holding on to the entire table (and associated memory) while parsing it

- Fixes
  - Tables now parsed one row at a time
  - Tokenizer cache shrunk after each block

⇨ lower memory usage + handle pathological pages

Overall progress (not just Q1 2014/15)

Performance

- Parsoid cluster load (last year)
- Parsoid cluster memory usage (last year)
- Pretty even & stable over last 6+ months

- Not on critical path for page loads
- Edited HTML goes through serialization

- Perf stats logged as part of RT testing, but nothing in production currently

Overall progress (not just Q1 2014/15)

Performance

Wiki	Metric	50th	95th	99th
ar	wt → html	0.27	4.40	21.0
	html → wt	0.10	0.45	1.0
en	wt → html	0.20	3.10	7.8
	html → wt	0.12	0.44	0.9
de	wt → html	0.26	2.77	6.6
	html → wt	0.11	0.41	0.8

(All values in seconds; RT testing, not prod data, consume with a pinch of salt, but trustworthy)

- Production html → wt likely faster than this because of selser. To be verified.
- With Parsoid HTML views, post-edit saves can be async and “instantaneous”.

Progress Q1:

Other

- Language variants:
  - Tokenizer fixes to handle variant syntax in place
  - Editing support work to be booted up
  - Work resumes in Q2
- Support for certain complex templates
  - A couple work-in-progress patches
  - Work resumes in Q2
- Wikimania 2014
  - First time for most of us

Progress Q1:

Wikitext linter GSoC project

- Wrapped up GSoC project (Hardik Juneja)
  - Flags wikitext “errors”
    - Missing end-tags, fostered content, etc.
  - LintTrap collects info during normal parsing
    - Sends it to LintBridge, an “external” webservice

- Future plan:
  - Identify more lint + collect stats
  - Enable in prod
  - Continue collaboration with Project CheckWiki
  - Community adoption
  - More work in later half of Q2

Progress Q1:

OCG-based PDF renderer

- Offline Content Generator (OCG) based
  - Parsoid HTML → LaTeX → PDF
- New renderer live as of Sep 29
- Old PDF renderer turned off

(Covered in OCG services review)

Things we didn’t get done in Q1

- Language variant editing
- Support for styling+content mixed templates (usually used in infoboxes)
- Yet to start work on stable element ids
- Work on native Gallery extension
  - Might get code from Wikia (Inez)
  - Encouraged Inez to submit a DOM spec as well
- Logs yet to be sent to LogStash
- Research: content widgets

Q2 tasks: Broad areas

- Parsoid HTML page views
- Supporting clients
- New applications
- Ongoing maintenance
- Production performance instrumentation

Caveat: We probably won’t get everything done

Q2 tasks: Parsoid HTML views

- Page views: work towards a prototype
  - Send Parsoid logs to LogStash -- high priority
    - Identify crashers / broken pages sooner
  - Continue plugging gaps in rendering accuracy
    - Cite CSS
    - Mixed content-style templates used heavily in some infoboxes
    - Tidy vs. HTML5 tree builder incompatibilities
  - Integrate with RESTBase
  - Move data-* attrs to metadata storage + migration plan for clients

Q2 tasks: Supporting clients

- Language variant support
  - Finish up basic parsing / editing support
  - Critical for VE on Chinese WP (and 8 other wikis)
- HTML editing for template parameters
  - Identify and fix perf issues with HTML support
  - Work with VE team to enable it
- Others
  - Gallery, LST, and other extensions used in non-WP projects.
  - PST support for VE
  - HTML/Wikitext switching in VE

Q2 tasks: New applications

- Start work on stable id support
  - Enabler for other projects

- PhantomJS-based PDF rendering port
  - Print CSS stylesheet for Parsoid HTML

Q2 tasks: New applications

- LintTrap integration
  - Helps identify / fix broken wikitext rather than support those use cases
  - Collaborate with Project CheckWiki

- Research + experiments
  - HTML content templates
  - HTML content widgets

Q2 tasks: Perf + maintenance

- Performance instrumentation
- Migrate Parsoid cluster to node 0.10
- Addressing incoming bug reports
- Technical debt
  - Promises API in code
- Pay more attention to rt-testing infrastructure
  - Some rough edges, crankiness in clients & server
  - Possibly move out of mysql db
  - Unclog perfstats so we can catch perf regressions

Thank you!

https://www.mediawiki.org/wiki/Parsoid

Backup slides

Stable ids: what?

- Parsoid parses page P with WT foo

  foo
- I edit source of P to bar\n\nfoo
- Parsoid reparses P:

  bar foo
- Can we re-assign id 1 to the second para?

Stable ids: why?

- Lets us associate metadata with content elements
  - authorship maps
  - efficient diffs
  - inline comments
  - content translation tracking
  - awesome features we haven’t thought about yet
- Lets us slim down HTML size, but still support switching between HTML & wikitext in VE