# Parsoid

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

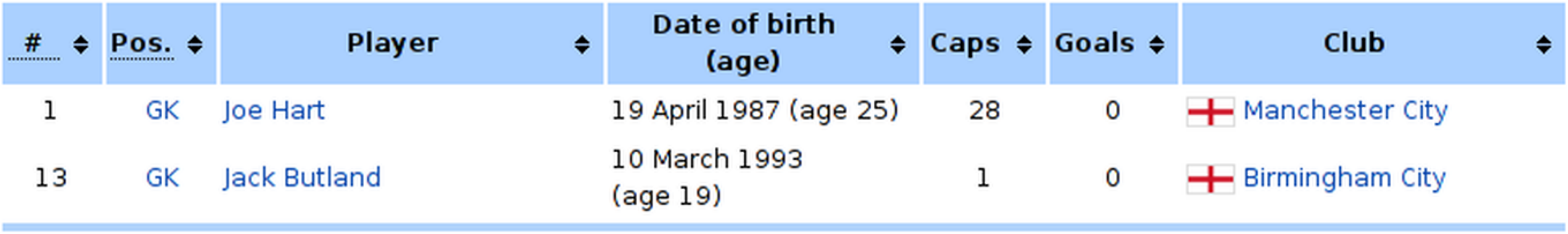
# Our objectives

Make it easy and efficient to view, reuse, and edit content.

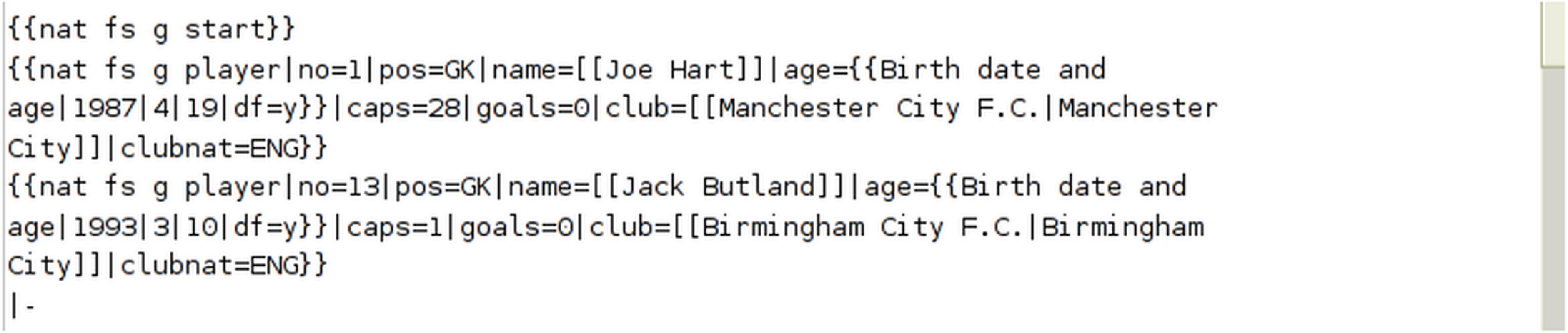
1. Convert wikitext content faithfully to semantic HTML+RDFa
2. Support HTML editing without dirty wikitext diffs
3. Use HTML in MediaWiki core
4. Support wikitext editing with Parsoid

# Parsoid's challenges

- Awesomeness of wikitext
  - fragments: {{table start}}{{table row}}{{table end}}



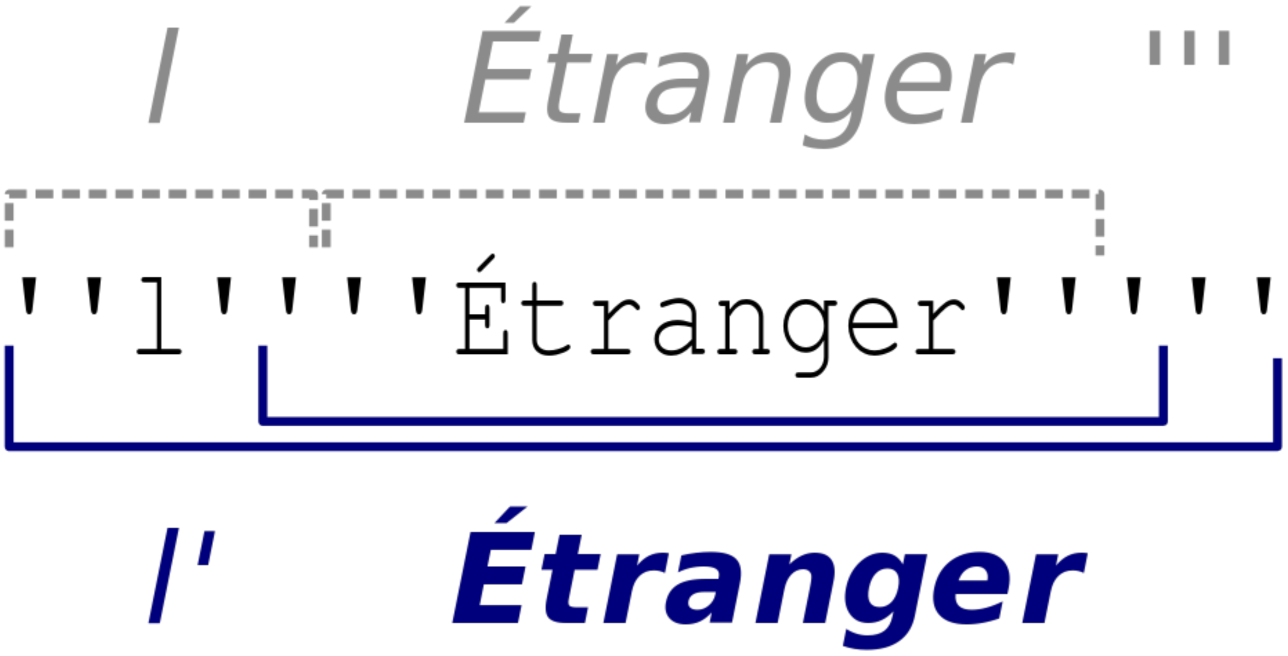
| # | Pos. | Player | Date of birth (age) | Caps | Goals | Club |
|---|---|---|---|---|---|---|
| 1 | GK | Joe Hart | 19 April 1987 (age 25) | 28 | 0 | Manchester City |
| 13 | GK | Jack Butland | 10 March 1993 (age 19) | 1 | 0 | Birmingham City |



```
{{nat fs g start}}
{{nat fs g player|no=1|pos=GK|name=[[Joe Hart]]|age={{Birth date and
age|1987|4|19|df=y}}|caps=28|goals=0|club=[[Manchester City F.C.|Manchester
City]]|clubnat=ENG}}
{{nat fs g player|no=13|pos=GK|name=[[Jack Butland]]|age={{Birth date and
age|1993|3|10|df=y}}|caps=1|goals=0|club=[[Birmingham City F.C.|Birmingham
City]]|clubnat=ENG}}
|-
```

# Parsoid's challenges

- Awesomeness of wikitext
  - fragments: {{table start}}{{table row}}{{table end}}
  - Context sensitivity
    - overlaps: <b>bold <i>and italic</b> only italic</i>
    - Quote balancing

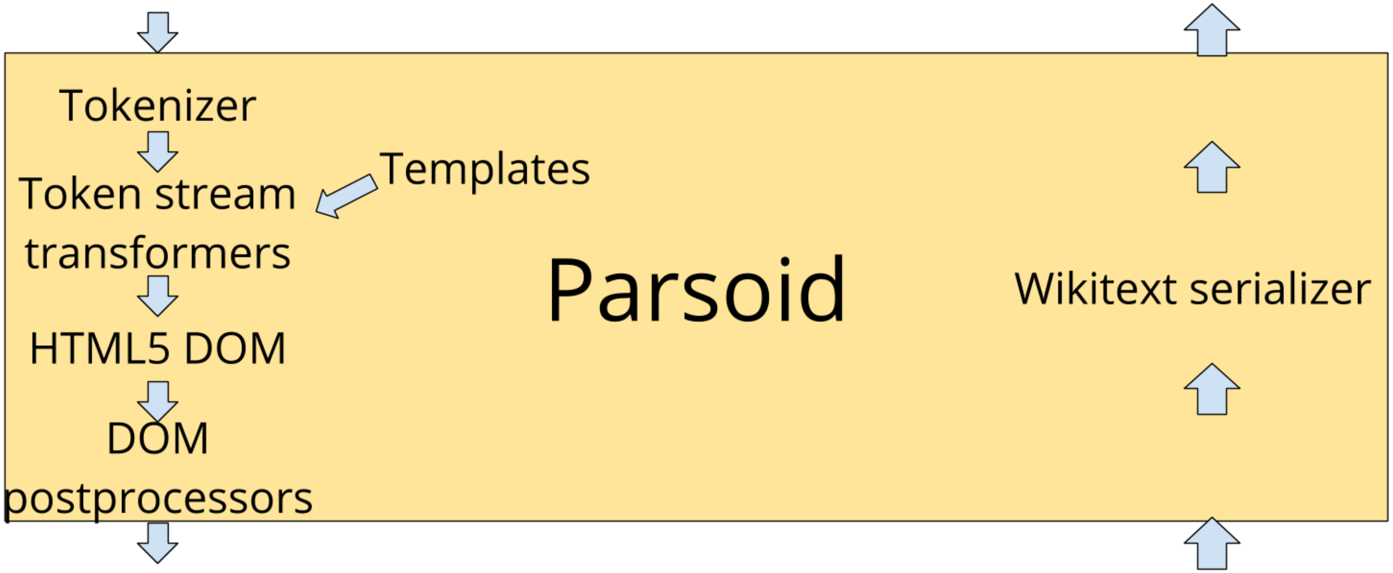


# Parsoid's challenges

- Awesomeness of wikitext
  - fragments: {{table start}}{{table row}}{{table end}}
  - Context sensitivity
    - overlaps: <b>bold <i>and italic</b> only italic</i>
    - `’’l’’’’Étranger’’’’’`
  - Lots of PHP parser bugs and irregularities
- Round-tripping
  - support wikitext storage and editing, **no dirty diffs**
  - [[Foo|bar]] , {{echo|[[Foo|bar]]}}, [[{{echo|Foo}}|bar]], [[... ]] .. all parse to the same html
- Performance
  - need to preserve much more information

**Wikitext**

```
'''''The Stranger''''' (''[[French language|French]]'':
''l'<nowiki />'''Étranger''''') is a [[novel]] by [[Albert
Camus|Camus]].
```



**HTML / RDFa**

```
<b><i>The Stranger</i></b> (<i><a rel="mw:WikiLink"
href="./French_language">French</a></i>: <i>l'<b>Étranger</b>
</i>) is a <a rel="mw:WikiLink" href="./Novel">novel</a> by
<a rel="mw:WikiLink" href="./Albert Camus">Camus</a>.
```

HTTP

**Browser**

***The Stranger*** (*French*: *l'**Étranger***) is a novel by Camus.

**VisualEditor**

**Bots, caches etc.**

# Digesting wikitext

- PEG tokenizer (unlimited look-ahead)
- context sensitive transformations on token stream, results in HTML tokens (start / end tag, text, comments etc)
- HTML DOM tree building fixes up nesting issues
- DOM post-processing establishes transclusion-affected DOM ranges, DOM source ranges, detects auto-inserted start / end tags

# Round-tripping wikitext

- Selective serialization based on DOM source range (DSR)
  - Use original wikitext for unmodified parts of the page

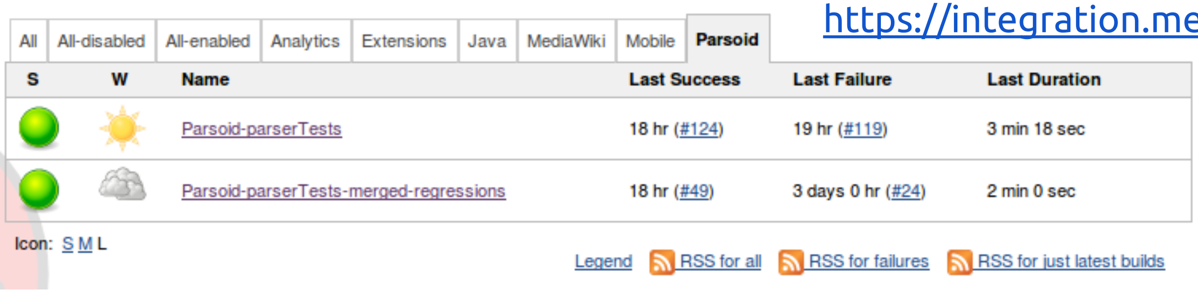
```
data-parsoid='{"dsr":[619,1264,0,0]}'
```

- Round-trip information in data-parsoid attribute

# Parsoid testing

CI with Jenkins: 1128 wikitext / HTML pairs, used in 5 modes for a total of 15068 tests

https://integration.mediawiki.org/ci/view/Parsoid



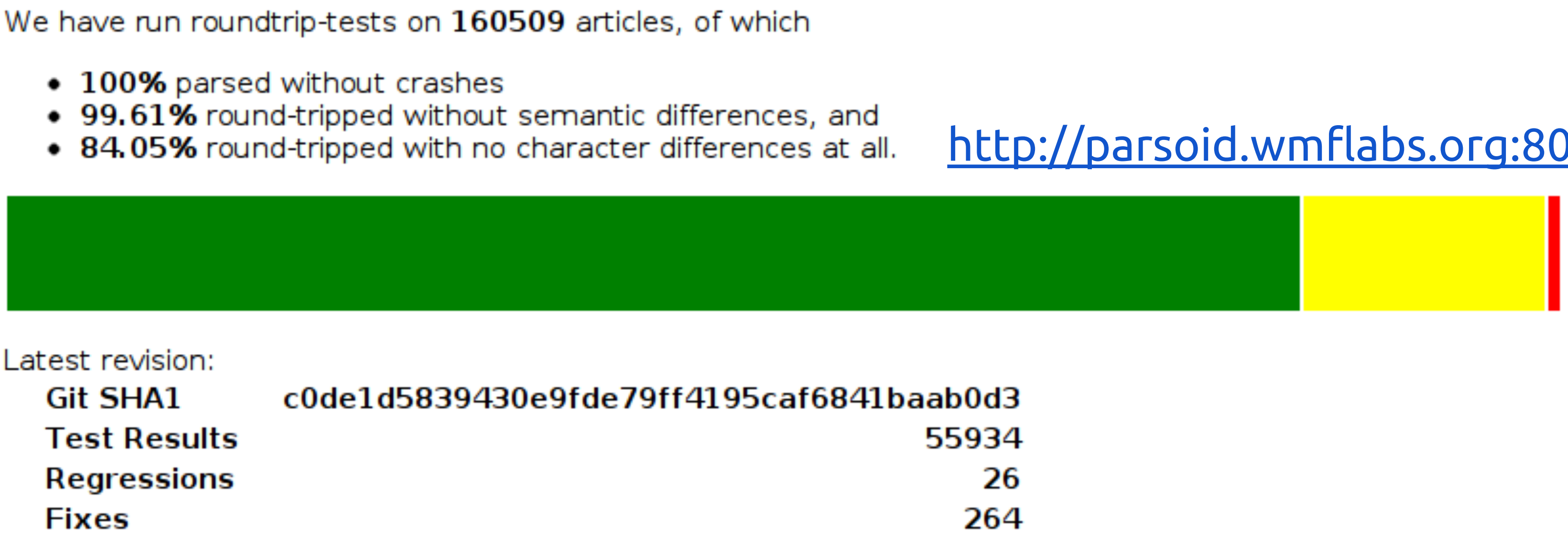
All | All-disabled | All-enabled | Analytics | Extensions | Java | MediaWiki | Mobile | **Parsoid**

| S | W | Name | Last Success | Last Failure | Last Duration |
|---|---|---|---|---|---|
| ● | ☀ | Parsoid-parserTests | 18 hr (#124) | 19 hr (#119) | 3 min 18 sec |
| ● | ☁ | Parsoid-parserTests-merged-regressions | 18 hr (#49) | 3 days 0 hr (#24) | 2 min 0 sec |

Icon: S M L

Legend  RSS for all  RSS for failures  RSS for just latest builds

Round-trip testing on 160k pages from 16 Wikipedias



We have run roundtrip-tests on **160509** articles, of which

- **100%** parsed without crashes
- **99.61%** round-tripped without semantic differences, and
- **84.05%** round-tripped with no character differences at all.

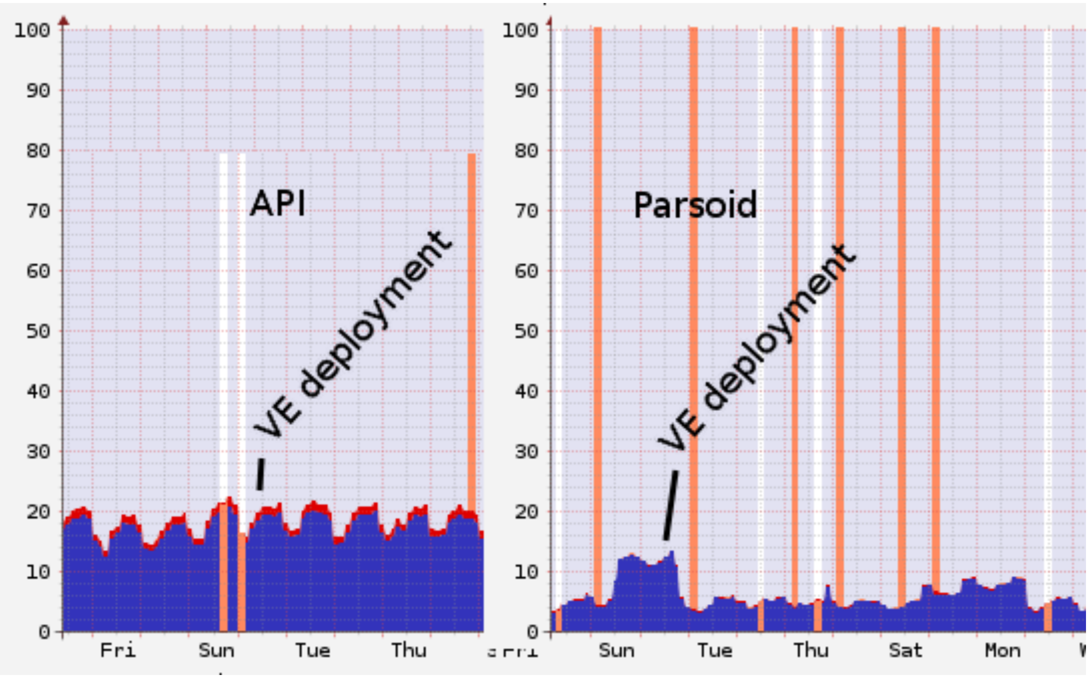
http://parsoid.wmflabs.org:8001/

Latest revision:

| | |
|---|---|
| **Git SHA1** | c0de1d5839430e9fde79ff4195caf6841baab0d3 |
| **Test Results** | 55934 |
| **Regressions** | 26 |
| **Fixes** | 264 |

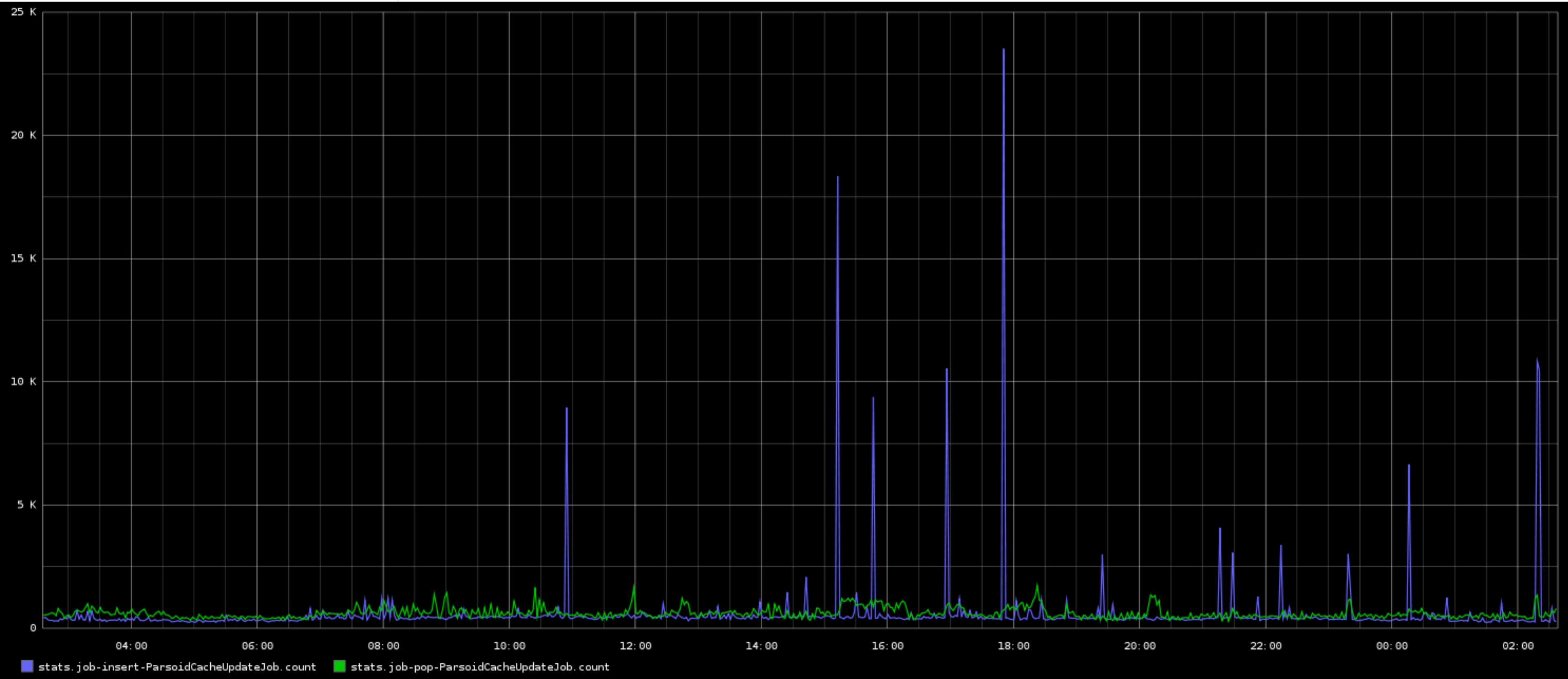
# Performance and scaling

- Async Node.js implementation running on 24 machines
- Caching setup, on-edit parsing and expansion reuse



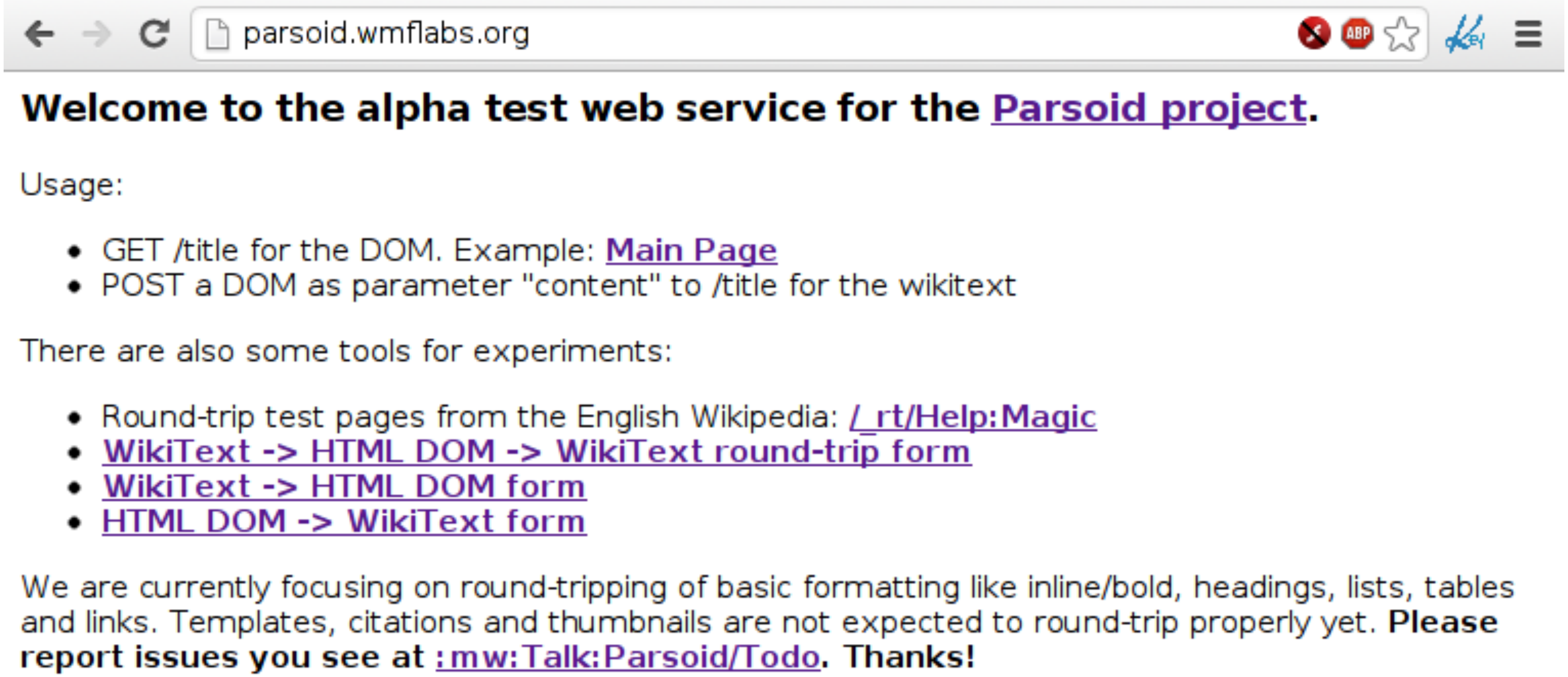
# Performance and scaling

- Async Node.js implementation running on 24 machines
- Caching setup, on-edit parsing and



# Try it!

http://parsoid.wmflabs.org (MW API soon)



**Welcome to the alpha test web service for the Parsoid project.**

Usage:

- GET /title for the DOM. Example: Main Page
- POST a DOM as parameter "content" to /title for the wikitext

There are also some tools for experiments:

- Round-trip test pages from the English Wikipedia: /_rt/Help:Magic
- WikiText -> HTML DOM -> WikiText round-trip form
- WikiText -> HTML DOM form
- HTML DOM -> WikiText form

We are currently focusing on round-tripping of basic formatting like inline/bold, headings, lists, tables and links. Templates, citations and thumbnails are not expected to round-trip properly yet. **Please report issues you see at :mw:Talk:Parsoid/Todo. Thanks!**

# Goals for 2013/14

- Continue work editing support and bug fixes
- Start to leverage HTML in MediaWiki core
  - HTML and page property storage (also for Flow?)
  - HTML diffing, basic authorship maps
  - Parsoid HTML for page views (might take longer)
- Investigate HTML-based templating
- Later: Switch MediaWiki storage to HTML, use Parsoid for wikitext editing

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

# Questions?

**More:**

**http://www.mediawiki.org/wiki/Parsoid**

# Tasks Q3 2013

- Image editing refinements [straightforward]
- Provide public HTML API [straightforward]
- Research: Language variant support [hard, Q3-Q4?]
- Research: Support switching between HTML and Wikitext within one edit [hard, Q3-Q4 2013?]

# Tasks Q3 2013

- HTML / Wikitext compound storage; support Flow [medium, Q3-Q4 2013]
- Enforce proper nesting of transclusions [hard, Q3-Q4 2013]
- Testing infrastructure improvements [straightforward, Q3-Q4 2013]
- Performance: More efficient template updates [straightforward]

# Other tasks on the horizon

- Parse most transclusion parameters to DOM once type info is available [medium, likely Q4 2013]
- HTML-only wiki support [hard, Q4 2013 - Q2 2014]
- Non-Wikipedia projects [likely hard, Q4 2013 - Q1 2014?]
- Research DOM-based templating [hard, Q4 2013 - Q1 2014]
- Use Parsoid HTML for all page views [hard, stretch goal, Q2 2014?]