Wikitext:

Upcoming changes, Available tools, How you can help

S. Subramanya Sastry, James D. Forrester

Wikimania 2017

https://etherpad.wikimedia.org/p/Wikimania2017-wikitext-aug11

First things first

Wikitext “parsers”

PHP Parser: default parser (since 2003)

Desktop view, Mobile web, iOS app, Action API

WIKIMEDIA
FOUNDATION

Wikitext “parsers”

Parsoid: alternate parser (since 2012)

VisualEditor, 2017 Wikitext Editor, Flow, Content Translation, Android App, Linter Extension, Kiwix Offline Reader, Google, REST API

WIKIMEDIA FOUNDATION

Wikitext “parsers”

WIKIMEDIA
FOUNDATION

Parsing Team @ WMF

Input

Advance wikitext as a language

Easier to write, faster to parse, less error-prone

WIKIMEDIA
FOUNDATION

Make wikitext content easier to analyze

Expose wikitext semantics, no parsing required

Output

WIKIMEDIA FOUNDATION

“Parsers”

Unify parsers: Same parser for reads as well as edits

WIKIMEDIA
FOUNDATION

Parsing Team Mission

- **Input**: Advance wikitext as a language
  - Easier to write, faster to parse, less error prone
- **Output**: Make wikitext content easier to analyze
  - Expose wikitext semantics in well-specified output
- **Parsers**: Unify parsers
  - Same parser for reads as well as edits

WIKIMEDIA
FOUNDATION

WIKIMEDIA
FOUNDATION

Why should you care?

- **Reasoning:** (Structured semantics for wikitext)
  - Improved ability to reason about wikitext (edits, templates)
- **Performance:** (Incremental parsing for high-performance edits)
  - Edits to popular templates won't hurt performance
- **Edits:** (Finer granularity edits → fewer edit conflicts)
  - List items, table cells, quoted text, html tags possibly
  - "Safer" templates, "Easier" template arguments

How this will affect you

- Changes to wikitext (affects editors)
  - Fix edge cases, improve semantics, make slow incremental changes
- Changes to HTML (affects gadget authors, HTML clients)
  - HTML5, semantic markup with a versioned spec
- Changes to parser hooks (affects extension authors)
  - Parser internals not exposed to extensions

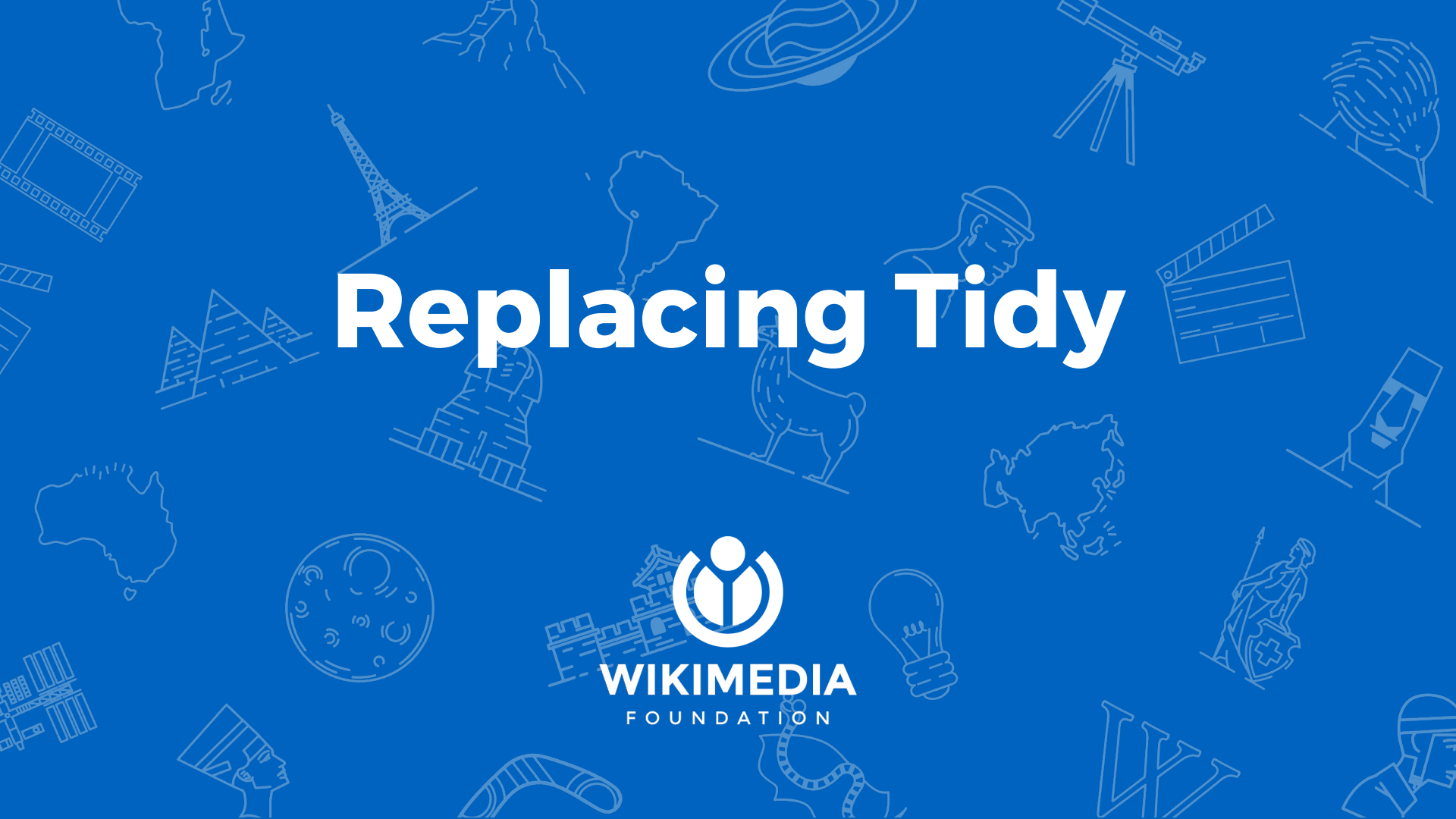
Specifics, please!

Examples

- **Finished**
  - Preprocessor changes to fix edge cases in language converter code
- **Ongoing**
  - Replacing Tidy (HTML4) with RemexHTML (HTML5)
  - Identifying broken markup via Linter extension
- **Upcoming**
  - Rendering images with <figure> instead of **(2017)**
  - New heredoc-style syntax for multi-template blocks **(2017–2018)**
  - “Balanced” templates **(2018)**

WIKIMEDIA
FOUNDATION

Replacing Tidy



HTML4 → HTML5

- HTML5-based library (RemexHTML) will replace Tidy
- Please read the FAQ on wiki

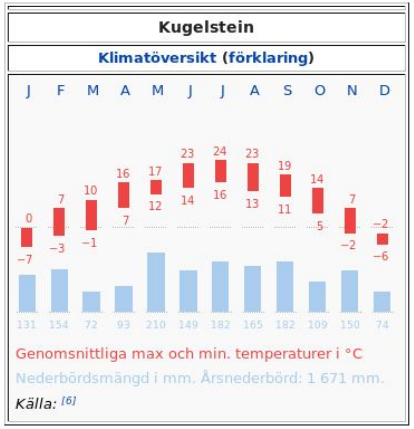
Some wikitext might display differently!

Especially broken wikitext / HTML

Or pages that rely on Tidy bugs

WIKIMEDIA
FOUNDATION

och den torraste är mars, med 72 mm nederbörd.[7]



Kommentarer

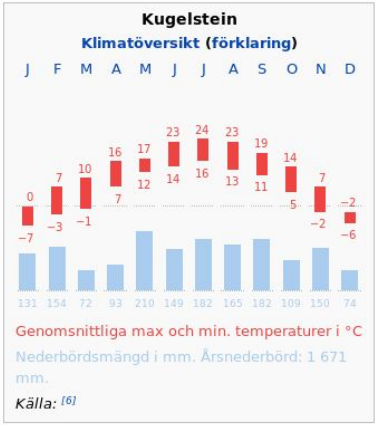
a. ^ Framräknat ur variansen i alla höjduppgifter (DEM 3") från Viewfinder Panoramas, inom 10 km radie.[2] Mer om algoritmen finns här: Användare:Lsjbot/Algoritmer.
b. ^ Den punkt som syns högst över den lokala horisonten runt platsen, enligt höjduppgifter i GeoNames.[1]
c. ^ Signifikant med genomsnittlig förekomst av namngivna sådana på jorden, enligt GeoNames.[1]

Källor

1. ^ [a b c d] Kugelstein at GeoNames.Org (cc-by); post uppdaterad 2013-07-28; databasdump nerladdad 2015-12-01
2. ^ "Viewfinder Panoramas Digital elevation Model". Läst 21 juni 2015.
3. ^ "NASA Earth Observations: Population Density". NASA/SEDAC. Läst 30 januari 2016.

Tidy

och den torraste är mars, med 72 mm nederbörd.[7]



Kommentarer

a. ^ Framräknat ur variansen i alla höjduppgifter (DEM 3") från Viewfinder Panoramas, inom 10 km radie.[2] Mer om algoritmen finns här: Användare:Lsjbot/Algoritmer.
b. ^ Den punkt som syns högst över den lokala horisonten runt platsen, enligt höjduppgifter i GeoNames.[1]
c. ^ Signifikant fler inom 20 km radie jämfört med genomsnittlig förekomst av namngivna sådana på jorden, enligt GeoNames.[1]

Källor

1. ^ [a b c d] Kugelstein at GeoNames.Org (cc-by); post uppdaterad 2013-07-28; databasdump nerladdad 2015-12-01
2. ^ "Viewfinder Panoramas Digital elevation Model". Läst 21 juni 2015.
3. ^ "NASA Earth Observations: Population Density". NASA/SEDAC. Läst 30 januari 2016.
4. ^ "NASA Earth Observations: Land Cover Classification". NASA/MODIS. Läst 30 januari 2016.
5. ^ Peel, M C; Finlayson, B L; McMahon, T A (2007). "Updated world map of the Köppen-Geiger climate classification". *Hydrology and* 544. doi:10.5194/hess-

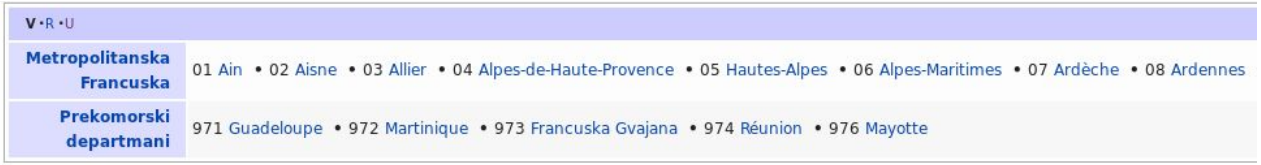
RemexHTML

Set Index". NASA. Läst 30

1 month - TRMM)".
NASA/Tropical Rainfall Monitoring Mission. Läst 30 januari 2016.

Snapshotted via ?action=parsermigration-edit

WIKIMEDIA FOUNDATION



Tidy

RemexHTML

Snapshotted via ?action=parsermigration-edit

WIKIMEDIA FOUNDATION

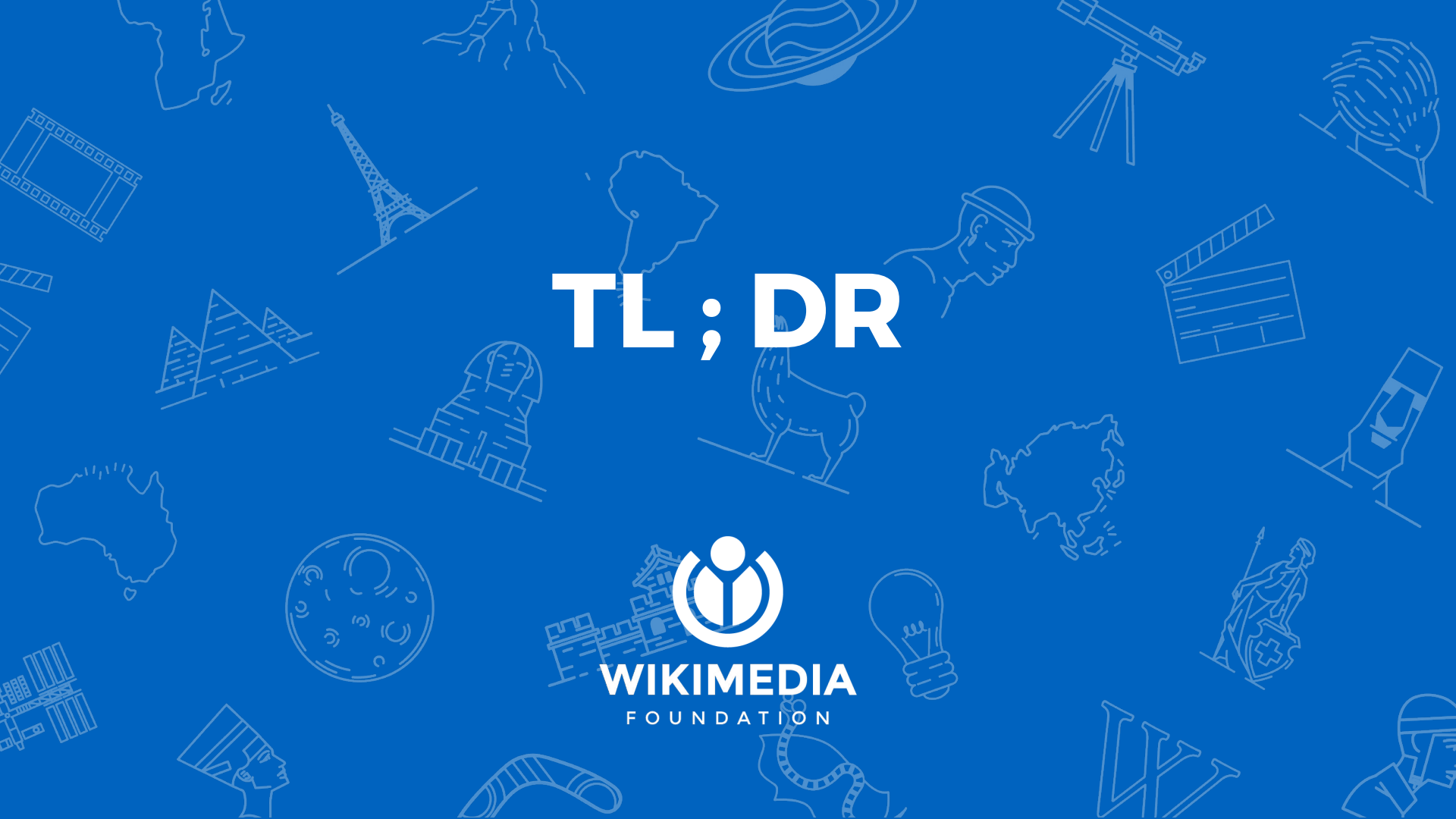
How you can help

- Fix High Priority errors exposed by Extension:Linter
  - Exposed via `Special:LintErrors` and the REST API
  - Extension:ParserMigration will help you verify fixes
    - `?action=parsermigration-edit` instead of `?action=edit`
- Status tracking
  - Dashboard (https://tools.wmflabs.org/wikitext-deprecation/)
  - Weekly stats published at mw:Parsing/Replacing_Tidy/Linter/Stats
  - Weekly pixel-diff run results at http://mw-expt-tests.wmflabs.org/

WIKIMEDIA
FOUNDATION

Pixel diff example (Tidy vs RemexHTML)

TL ; DR



Summary: Get Involved

- Incremental changes to wikitext, HTML, parser APIs
  - Pages, templates, bots, gadgets, extensions might need fixing
- Some changes:
  - Tidy (HTML4) → RemexHTML (HTML5) (FAQ)
  - <figure> markup for images (T118517)
  - Improved syntax for multi-template content blocks (T114432)
  - Balanced templates (T114445)
- **Tools:** Linter, ParserMigration, Pixel diffs, Dump greps

https://www.mediawiki.org/wiki/Parsing/Get_involved

WIKIMEDIA
FOUNDATION

DISCUSS!
THANK YOU!

Backup Slides

Heredoc syntax (not final)

Now: Four “unconnected” transclusions

{{tablestart}}

{{tablerow|1}}

{{tablerow|2}}

{{tableend}}

One Proposal: Single “connected” transclusion

{{<<MyTable}}

{{tablerow|1}}

{{tablerow|2}}

{{>>MyTable}}

WIKIMEDIA FOUNDATION

Benefits

- Eliminates escaping hazards by hoisting content out of template argument position
- Supports balanced templates goal

Balanced Templates

- Add (type) annotation to templates indicating use context

  `{{#balance}}, {{#balance:inline}}, {{#balance:block}}, {{#balance:attribute}}, {{#balance:string}}` ....
- Type enforced by parsing to:
  - Well-formed & balanced DOM → unclosed tags closed
  - Attribute → Appropriate escaping without explicit nowikis
  - String → Appropriate escaping without explicit nowikis
  - ....

Benefits

- Predictable rendering across all uses
- Markup errors are constrained to limited context (table, list, paragraph, etc.)
- Enables higher performance via incremental parsing