Future of Structured Documents

VisualEditor, Citations, and Wikidata — Oh my!

2015-07-18
Wikimanía 2015
Ciudad de México

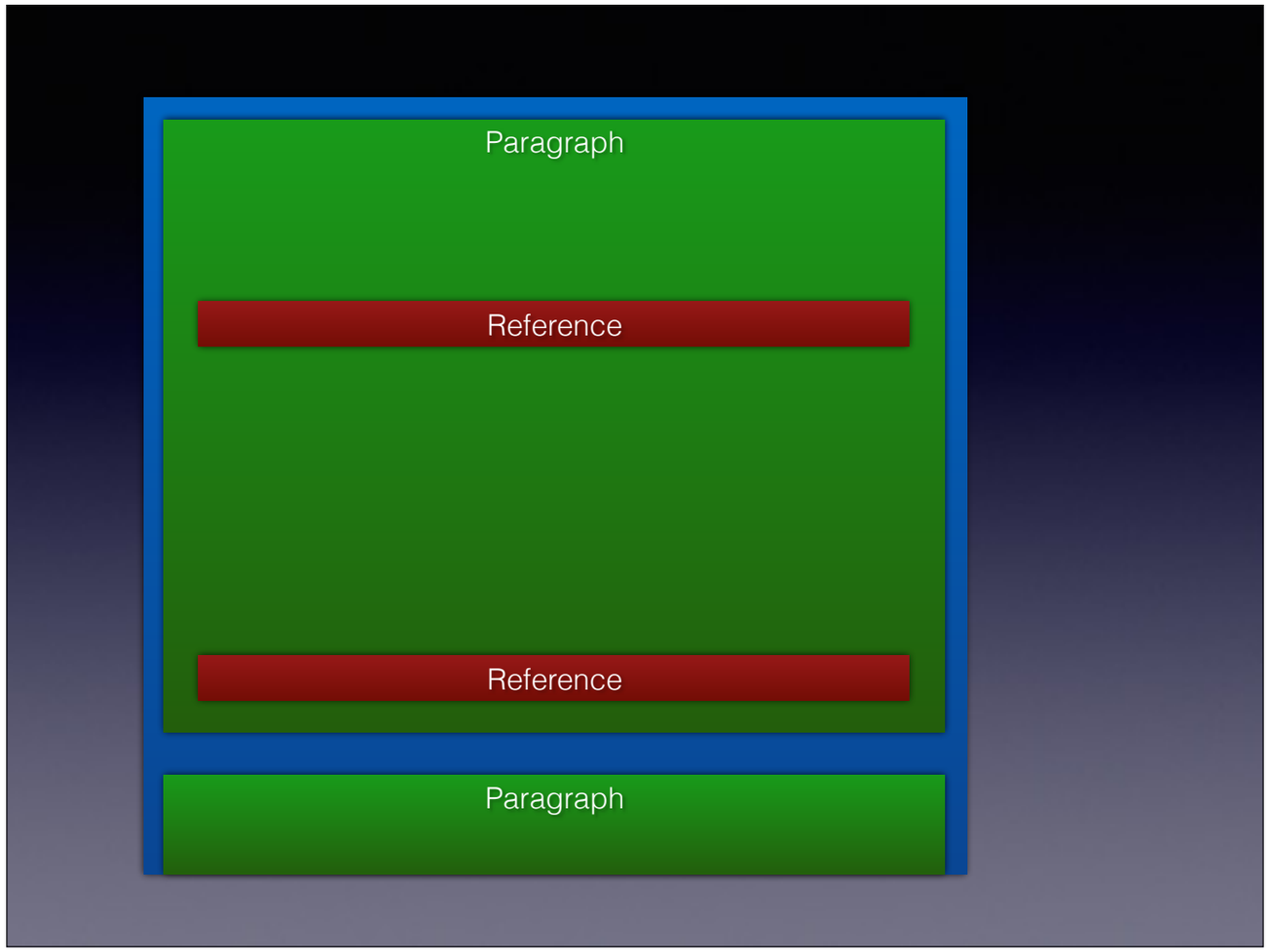
Brion Vibber
Software Architect
Wikimedia Foundation

Today: BIG BLOB OF TEXT

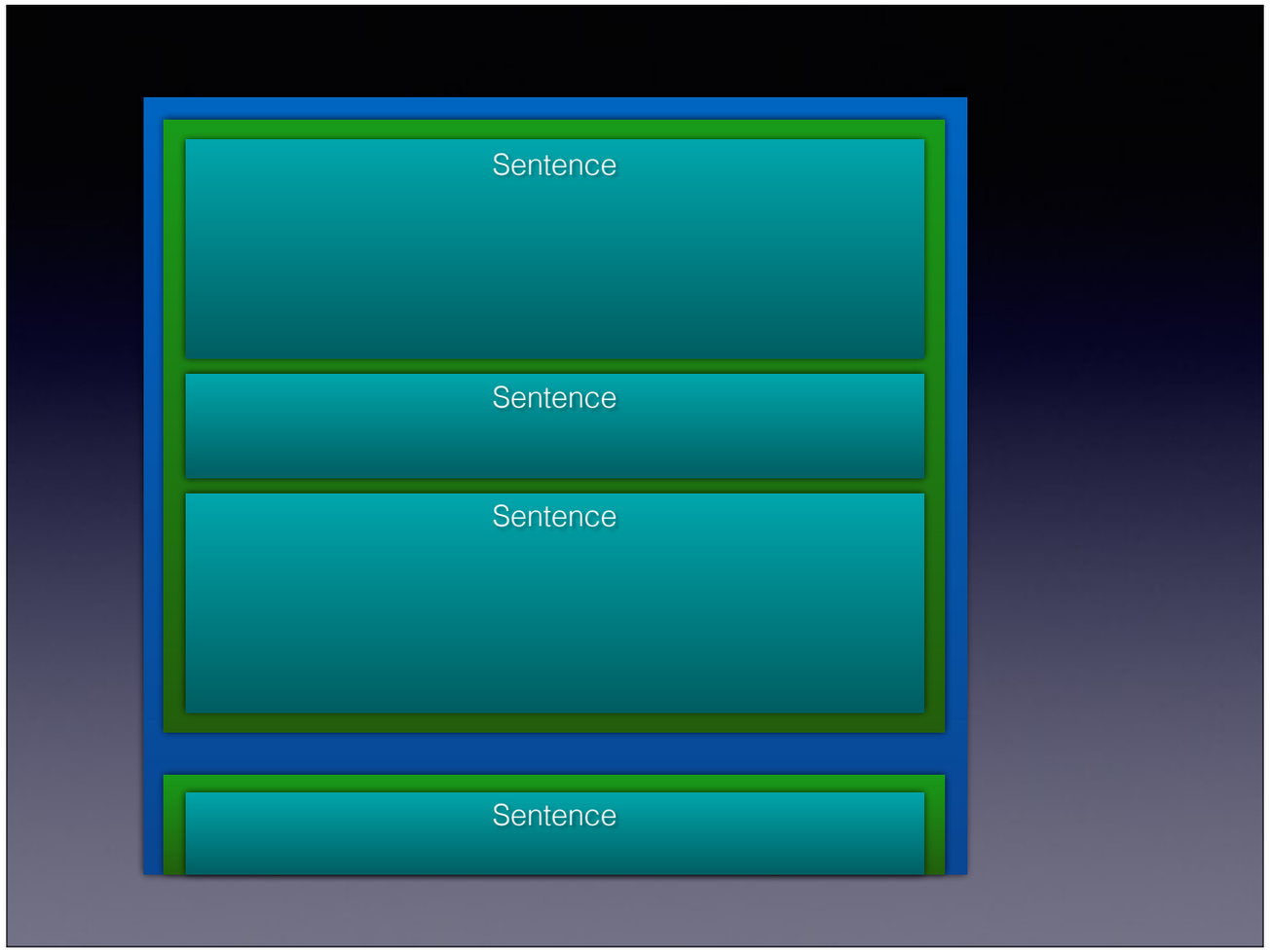
'''Nahuatl''' ({{IPAc-en|'|n|ɑː|w|ɑː|t|əl}};<ref>Laurie Bauer, 2007, ''The Linguistics Student's Handbook'', Edinburgh</ref> {{IPA-nah|'naːwatɬ|-|nawatl.ogg}}<ref group=cn>The Classical Nahuatl word ''{{lang|nah|nāhuatl}}'' (noun stem '''''nāhua''''', + absolutive '''''-tl''''') is thought to mean "a good, clear sound" {{harvcoltxt|Andrews|2003|pages=578,364,398}} This language name has several spellings, among them náhuatl (the standard spelling in the Spanish language), ({{cite web |url=http://rae.es/drae/?type=3&val=n%C3%A1huatl|title=Náhuatl |publisher=rae.es |language=Spanish |accessdate=6 July 2012 }}) Naoatl, Nauatl, Nahuatl, Nawatl. In a back formation from the name of the language, the ethnic group of Nahuatl speakers are called ''Nahua''.</ref>), known informally as '''Aztec''',<ref name="Glottolog" /> is a language or group of languages of the [[Uto-Aztecan]] [[language family]]. Varieties of Nahuatl are spoken by an estimated {{nowrap|1.5 million}} [[Nahua peoples|Nahua people]], most of whom live in Central [[Mexico]]. All [[Nahuan languages]] are indigenous to [[Mesoamerica]].

Nahuatl has been spoken in [[Central Mexico]] since at least the 7th

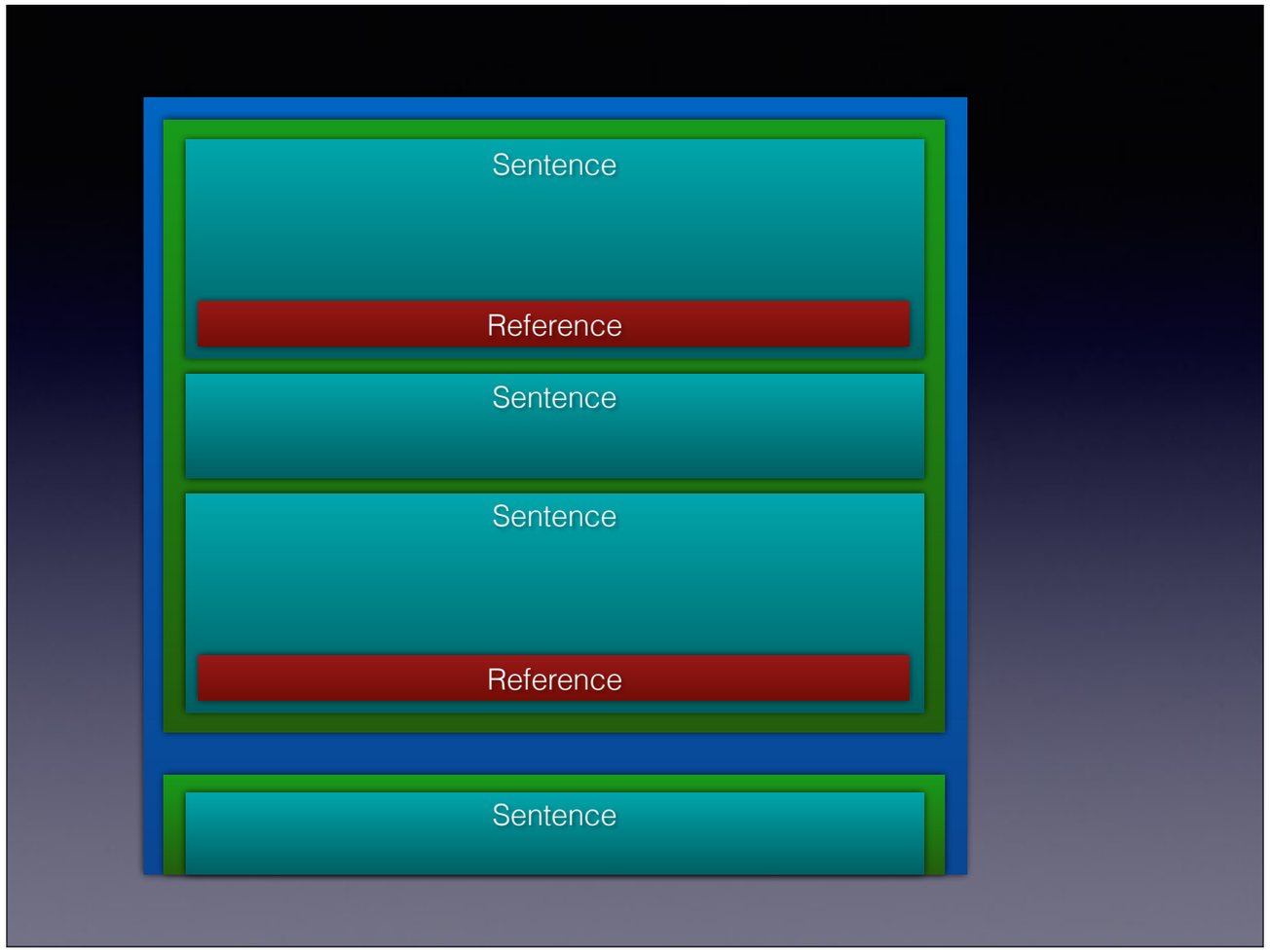
Eek scary!



Our current data model is pretty primitive. Paragraphs occasionally contain references indicating a source of information, but they have unclear scope and semantics.



What if we actually divided up paragraphs into smaller addressable sentence chunks?



We could then scope a contained reference to the text of the sentence.

Benefits?

- **Addressable content**
  - Commonality and cross-referencing
  - identity preserved across edits
- Cleaner extracts
  - Make it easier to jump around large amounts of text
  - Search for specific types of claims within an article?

Does the same claim get used in multiple articles? In multiple languages? In a diagram or video presentation?
Sentences identity preserved across edits, even moves between paragraphs or pages or languages or media!
Text can be more cleanly extracted and used.

The horror of wikitext

- Imagine wrapping every sentence in a tag or parser function…
  - {{#sentence:63d70d24a9|[[Mexico]] is a pretty awesome place.<ref>…</ref>}}

Imagine a world, where every sentence is wrapped in a parser function or tag to specify its id. Nobody wants to edit that in wikitext!

VisualEditor to the rescue?

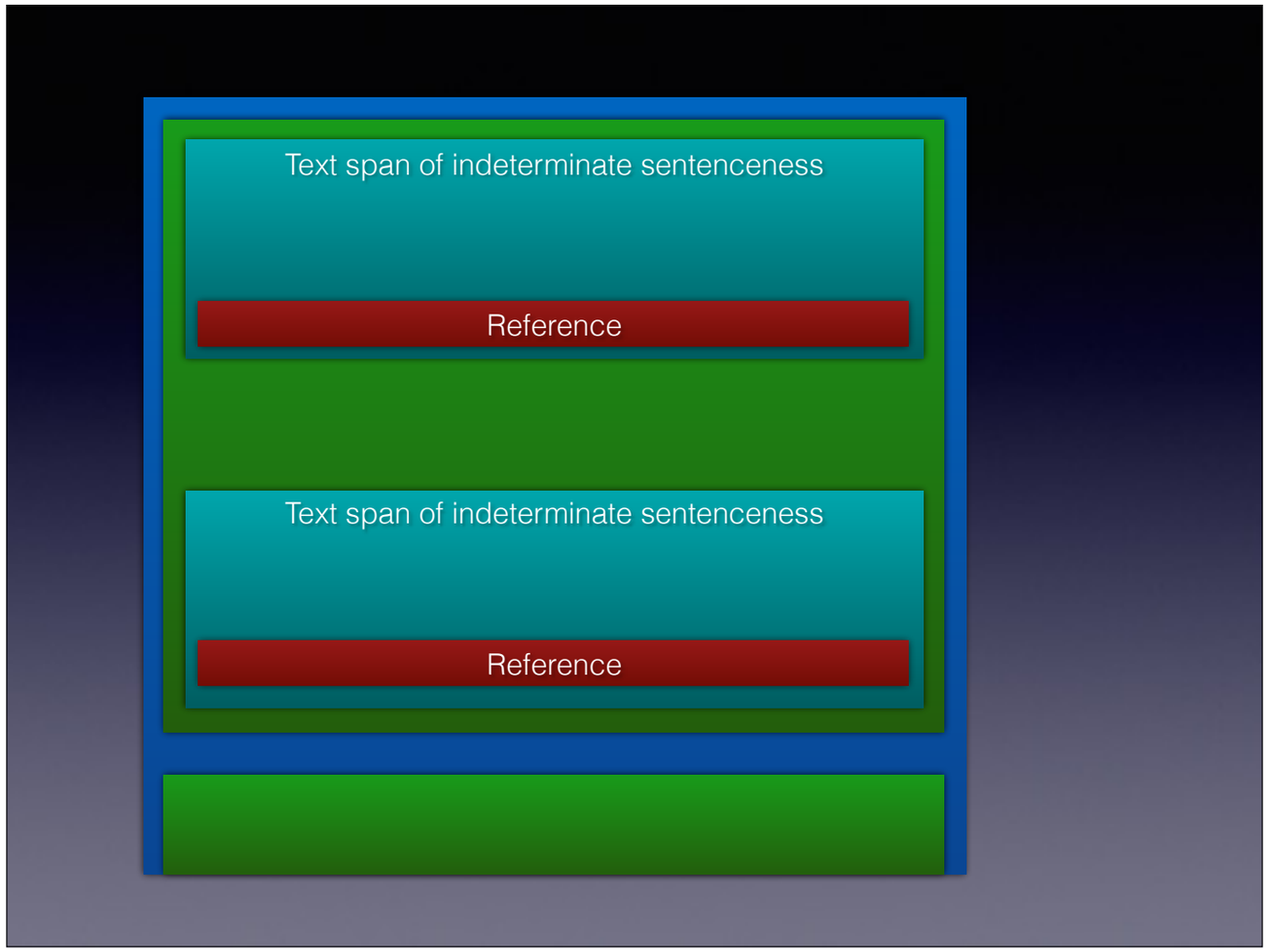
- A -like wrapper around a sentence would be invisible and unobtrusive, even if it comes with a big scary ID number:
  - **Mexico** is a pretty awesome place![1]

Considerations

- complex sentences may combine multiple claims
- many claims may be hard to define a data point for
- may need visual hints to show boundaries
- what about combined/split sentences?

Let’s backtrack a little

- ~~all sentences~~ -> **referenced text spans**



Instead of trying to generalize to every bit of information, let’s consider an opt-in addressability scheme: allow scoped text spans to be added around references, but don’t require it on everything. Boundary relations will be clearer, and unreferenced sentences don’t need to take on the extra burden.

What was that about Wikidata?

Oh right! That’s the fun part.

Let’s move away from the addressability issue a bit and look at the references themselves.

Today: SMALL BLOBS OF TEXT

```
<ref>Laurie Bauer, 2007, ''The Linguistics Student's Handbook'', Edinburgh</ref>

<ref group=cn>The Classical Nahuatl word ''{{lang|nah|nāhuatl}}'' (noun stem '''''nāhua''''', + absolutive '''''-tl''''' ) is thought to mean "a good, clear sound" {{harvcoltxt|Andrews|2003|pages=578,364,398}} This language name has several spellings, among them náhuatl (the standard spelling in the Spanish language),({{cite web |url=http://rae.es/drae/?type=3&val=n%C3%A1huatl|title=Náhuatl |publisher=rae.es |language=Spanish |accessdate=6 July 2012 }}) Naoatl, Nauatl, Nahuatl, Nawatl. In a back formation from the name of the language, the ethnic group of Nahuatl speakers are called ''Nahua''.</ref>

<ref name="Suárez 1983:149">{{harvcoltxt|Suárez|1983|page=149}}</ref>
```

The actual reference contents themselves today are freeform wikitext, sometimes run through templates for common formatting.

Problems today

- Difficult to identify use of a particular resource
- No commonality between articles using the same reference (maybe a template if lucky)
- No commonality between languages/projects using the same reference

Potential

- References themselves can be given a consistent ID
- Can easily cross-reference all uses of a particular reference source; sounds useful for maintenance
- When translating or copying across projects, keep the same reference intact!
- Can provide a linkable URL for non-web or archived/offline resources

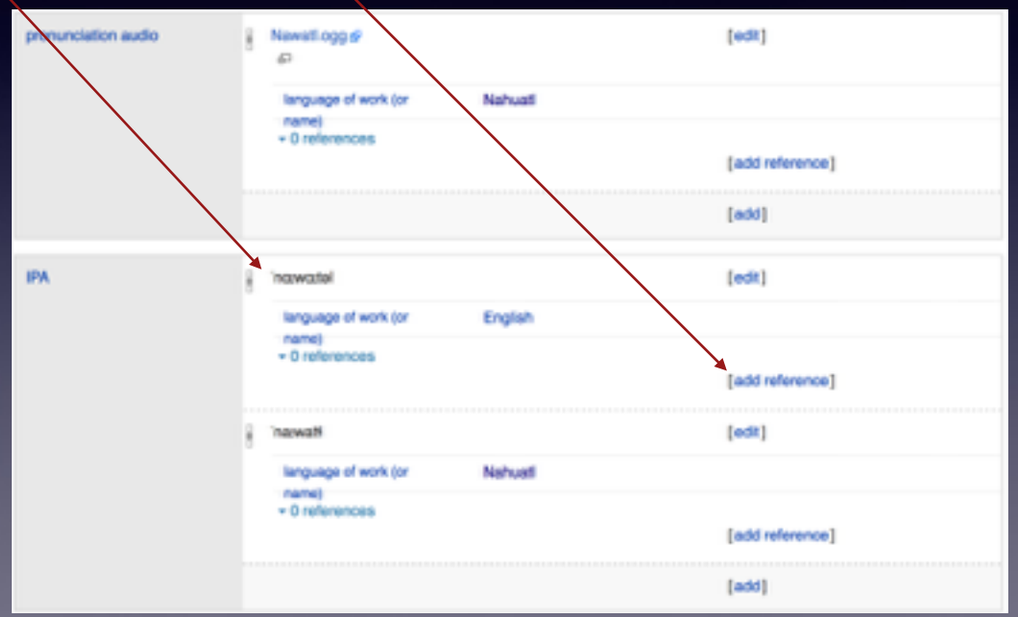
Potential

- Use the same references system in other media types?
  - PDF/ebook output
  - extend ‘spoken Wikipedia’ with links
  - clickable refs in videos

The data part?

That’s all well and good right? But Wikidata is more than a reference repository — it allows making machine-readable statements about things.

/ˈnɑːwɑːtəl/;[5]



What if Wikidata could hold the magic reference in the [5] as well as the data of the pronunciation?

It can’t be that easy?

- Wikidata references seem currently limited/ confusing?
- What about hard-to-define statements?

As a user, I found it hard to figure out how to use the reference field in wikidata statements… and some things are going to be complex or hard to define in machine-readable format.

Nahuatl has been spoken in [[Central Mexico]] since at least the 7th century AD.<ref name="Suárez 1983:149">{{harvcoltxt|Suárez|1983|page=149}}</ref>

Ok what can we do with this one? Hard to create a clear explanation; should probably not expect everyone who edits to be able to figure out how to encode ‘Nahuatl’ -> ‘is-spoken-in’ -> ‘Central Mexico’ (qualifier: time-since ‘7th centure’) as wikidata statements…

Ref data stubs?

- Start with entity, text, and reference
- Let the Wikidata community figure out how to encode the data details
- All Wikipedia articles using the reference will automatically get the attached data once done!

Maybe we should be able to create vague statements with a reference and a usage, and let the wisdom of the crowds narrow it into machine-readable information?

questions & ideas?

-end-