# Identifying and extracting data trapped in our content

## Session Setup - https://phabricator.wikimedia.org/T206073

Theme: Increasing our technical capabilities to achieve our strategy
Type:
Facilitation Exercise(s): Small group brainstorming/discussion

Leader(s): Michael Holloway
Facilitator: Kate
Scribe: Aaron

Description: Wikipedia and other Wikimedia projects contain lots of potentially structurable data that is represented in an unstructured way in wikitext or HTML. With new data capabilities being developed, we can now represent this data in a structured way. The purpose of this session is to identify data that is already present, but trapped in unstructured content, so we can explore techniques for modeling and extracting that data.

**Facilitator Instructions:** /Session_Guide#Session_Guidance_for_facilitators

## Questions to answer during this session

| **Question** | **Significance:**<br>**Why is this question important? What is blocked by it remaining unanswered?** |
|---|---|
| What types of data are currently stored in content that should be extracted and stored separately? What type of data is metadata and which is data to be composed into content?<br>(Specifically discuss Categories and | Identifying data within HTML content that we want extract into structured data is the first step in adding more semantic information about our content. This allows us to plan for the types of data that we want to store and design ways to extract the data. |

| | |
|---|---|
| Infoboxes.) | |
| Should the data you identified be stored on the host wiki or should it be stored on Wikidata? How do you decide this? | It is unclear where a lot of data should be stored and how we make this decision. Answering this allows us to plan where to store such data and provide future guidance to others. |
| Which types of data that were identified must support versioning? | Knowing which type of data must support versioning allows us to make decisions on how to store it and assess its impact on infrastructure. |
| Do you anticipate having difficulties automating the extraction of any of the data that you have identified? Do you anticipate having difficulties modeling any of the data that you have identified? Why? | Identifying data that has the potential for being difficult to extract or model will help us plan and prioritize extracting this data. |

## Attendees list

- Daren, Cindy, Gergo, Josh Minor, Kate, Aaron, Magnus, Subbu, Santosh, Michael Holloway, Ramsey, Jon Katz, Danny Horn, Lydia, Cheol, Marko

## Structured notes

There are five sections to the notes:

1. **Questions and answers:** Answers the questions of the session
2. **Features and goals:** What we should do based on the answers to the questions of this session
3. **Important decisions to make:** Decisions which block progress in this area
4. **Action items:** Next actions to take from this session
5. **New questions:** New questions revealed during this session

## Questions and answers

Please write in your original questions. If you came up with additional important questions that you answered, please also write them in. (Do not include “new” questions that you did not answer, instead add them to the new questions section)

# Identifying and extracting data trapped in our content

**Q**: What types of data are currently stored in content that should be extracted and stored separately? What type of data is metadata and which is data to be composed into content?

(Specifically discuss Categories and Infoboxes.)

**A** (If you were unable to answer the question, why not?):

- Infoboxes
- Categories -- subcategories
- See also
- ToC
- Quality assessment /warnings / stubs
- Navboxes (article relations)
- Template data
- Display title
- Template styles
- Workflow state (afd/afc/draft)
- Double underscore switches __NOINDEX__ __HIDDENCAT__
- Coordinates
- Tables/Lists
- Semantic statements as described in the text “portland has a population of”
- Spoken article
- Article series
- Proofread progress/index
- All of commons data
- Semantic mediawiki data

**Q**: Should the data you identified be stored on the host wiki or should it be stored on Wikidata? How do you decide this?

**A** (If you were unable to answer the question, why not?):

This varies by the kind of data, and might be influenced by social/political concerns (e.g., certain Wikipedias are not comfortable with displaying Wikidata content due to policy differences between the projects).

Also, it was pointed out that it’s not necessarily the case that these need to be *stored* in a structured way; rather, it might be sufficient to provide a structured *view* of the data (wherever stored). Need to separately

# Identifying and extracting data trapped in our content

consider data, logic, and presentation.

**Proposal: keep presentation local, and data and logic centralized.**

**Q**: Which types of data that were identified must support versioning?

**A** (If you were unable to answer the question, why not?):

Two answers were provided: either (a) all of them, and they already have an associated versioning scheme (namely, the revision ID of the page they came from); or (b) we don't know what "versioning" means here.

**Q**: Do you anticipate having difficulties automating the extraction of any of the data that you have identified? Do you anticipate having difficulties modeling any of the data that you have identified? Why?

**A** (If you were unable to answer the question, why not?):

No overarching difficulties were raised, but it was also clear that there's no single strategy for getting at all of this data.  That in itself is a difficulty.

## Features and goals

**For Use Case and Product Sessions:**
**Given your discussion of the topic and answering questions for this session, list one or more user stories or user facing features that we should strive to deliver**

1. Many different kinds of structurable data were identified but none were cited as especially urgent.  Which to pursue extracting and exposing via API will depend on product team needs.

# Identifying and extracting data trapped in our content

| | | |
|---|---|---|
| Why should we do this? | What is blocking it? | Who is responsible? |
| 2. | | |
| Why should we do this? | What is blocking it? | Who is responsible? |
| 3. | | |
| Why should we do this? | What is blocking it? | Who is responsible? |

## Important decisions to make

| **What are the most important decisions that need to be made regarding this topic?** | | |
|---|---|---|
| 1. None | | |
| Why is this important? | What is it blocking? | Who is responsible? |
| 2. | | |

# Identifying and extracting data trapped in our content

| | | |
|---|---|---|
| Why is this important? | What is it blocking? | Who is responsible? |
| 3. | | |
| Why is this important? | What is it blocking? | Who is responsible? |

## Action items

| **What action items should be taken next for this topic? For any unanswered questions, be sure to include an action item to move the process forward.** | | |
|---|---|---|
| 1. None | | |
| Why is this important? | What is it blocking? | Who is responsible? |
| 2. | | |
| Why is this important? | What is it blocking? | Who is responsible? |

# Identifying and extracting data trapped in our content

<table>
<tr><td colspan="3">3.</td></tr>
<tr><td>Why is this important?</td><td>What is it blocking?</td><td>Who is responsible?</td></tr>
</table>

## New Questions

<table>
<tr><th colspan="3">What new questions did you uncover while discussing this topic?</th></tr>
<tr><td colspan="3">1. How can Wikidata and the Wikipedias reach a compromise on policy so that it’s an acceptable “backend” for data displayed on Wikipedia when it’s most suitable on a technical level?</td></tr>
<tr><td>Why is this important?<br>Social/political concerns quickly arise when discussing integrating Wikidata content into other projects, or editing Wikidata from the context of other projects.  Solving this will improve Wikidata’s usefulness and ability to achieve its goals, and reduce the need for technical workarounds like the local title descriptions used in place of item descriptions from Wikidata on enwiki.</td><td>What is it blocking?<br>Full adoption of Wikidata data on the Wikipedias and other projects.</td><td>Who is responsible?<br>WMDE</td></tr>
<tr><td colspan="3">2.</td></tr>
<tr><td>Why is this important?</td><td>What is it blocking?</td><td>Who is responsible?</td></tr>
<tr><td colspan="3">3.</td></tr>
</table>

# Identifying and extracting data trapped in our content

| Why is this important? | What is it blocking? | Who is responsible? |
|---|---|---|
| | | |

## Detailed notes

Place detailed ongoing notes here. The secondary note-taker should focus on filling any [?] gaps the primary scribe misses, and writing the highlights into the structured sections above. This allows the topic-leader/facilitator to check on missing items/answers, and thus steer the discussion.

- E.g. Wiktionary wants to have every word in every language. Let's say we want to do something with that on Mobile. That's very hard because there's not really a consistent structure within a language wiktionary -- let alone between languages. E.g. MCS only works with English Wiktionary. *[Note: This is for illustrative purposes only, because Wikidata is fixing this particular problem.]*
- Brainstorming examples of data to extract (from whiteboard)
    - Infoboxes
    - Categories -- subcategories
    - See also
    - ToC
    - Quality assessment /warnings / stubs
    - Navboxes (article relations)
    - Template data
    - Display title
    - Template styles
    - Workflow state (afd/afc/draft)
    - Double underscore switches __NOINDEX__ __HIDDENCAT__
    - Coordinates
    - Tables/Lists
    - Semantic statements as described in the text "portland has a population of"
    - Spoken article
    - Article series
    - Proofread progress/index
    - All of commons data
    - Semantic mediawiki data
    - I'll visit the board and note this down.
- Magnus: Ideally, these things should exist on Wikidata.

  - If they are copied, they will get out of sync. So we need to display wikidata data on various Wikipedias. But on big wikis, there is social resistance. So we need to be careful where we store [this data].
- Cindy: what is the domain, do we include external wikis in this?
- Michael: no, not really while disclosing, but we should recognize that semantic mediawiki etc exists. [“our content” === the Wikimedia projects]
- Magnus: Depends on the wiki owner
- Aaron: is behavioral data (e.g. editor reputation based on content persistence measures) out of scope (yes)
- Magnus: Also lists. Lots of manually curated information. Lists should be re-creatable from Wikidata by a query.
- Josh M: Disambiguation pages!
  - Magnus: Yes. Could also go the opposite route.
- Josh M: New portal has a bunch of semantics buried. Timelines. Did you know facts.
  - Michael: Summarized as “main page content”
  - Gergo: Lists of frequently asked questions
- Darren: Relationships between pages in the form of links.
- Aaron: Cleanup templates! (In scope!)
  - SSmith: We have been working on a taxonomy for that.
- Darren: See relationships between content based on who contributes.
  - Magnus: Could cluster interests based on who works on what with a link graph
  - Josh M: ???
- Gergo: User boxes and other content on user pages.

Cutting off brainstorming. Break into groups.

- Should it be stored in Wikidata, the host wiki, or somewhere else.

The cool group: [https://etherpad.wikimedia.org/p/wmtc18_extracting_data](https://etherpad.wikimedia.org/p/wmtc18_extracting_data) (Star, Subbu, Halfak, pheudx, Ramsey)

DJ: Many of the problems can be solved by data, logic, presentation. Keep presentation local. Keep data and logic centralized.

- **Important points**:
- Extract can mean many different things (extract at source or delivery)
- Storing datat in structured way vs provide a strucutred view of data
- Do you need to recored the revision of the data from wikidata when put in a ppage
- What does “versioning” mean ?
- Differences in curation models have different social blockers

# Identifying and extracting data trapped in our content

- Many issue solvable by splitting data/logic/selection/presentation.

Group 1: Daren, Gergo, Lydia, Cindy, DJ

- **Q1**:
- Lydia: wiktionary. We started storing that in wikidata. So basically we already changed the location where we want to store it.
- What is the problem: structural,
- Some stuff doesn't belong in wikidata right.
- So decide on wghat need to be decentralized vs what needs to be extracted from wikitext. Two different things for different data.
- Catergories: MCR slot, structured done.
- Basically everything you put at the top or the bottom of an article is likely and MCR slot ?
- Wikidata: infobox partially, navboxes can be generated from , coordinates, tables (some)
- Centralized but not wikidata: user info
- MCR slot: categories,   page issues, proof read
- Article series
- Strategy: Store as wikitext in a MCR slot and stop bothering with it. (either hide there is an MCR slot or not
- **Q2**: What difficulties do you expect.
- Can we have seperate storage strategies for different things ?
- Give them a parserfunction for taking a piece of string and a keyname and extract that data towards an MCR slot.
- PPI try to store data that isn't actually structured or formatted… what to do with that.
- What we need is,
- So storage and curation of data in MCR and wikidata, but that doesn't necesarrily mean that we don't still have shadow indexed tables and apis to then reuse that information again.
- We need to recognize that not all encoded data needs the same approach towards extracting this information from the wikitext.
- Where does it belong (de)centralized (and does it need an override)
- Where is the canocnial information stored
- How is the information exposed.
- Which of these types of data need version
- **Q3**: Versioning:
- Non versioning: userpage info (maybe? If it's only private ?)
- Do you **locally snapshot revision information** if it is coming from an external or centralized location.

# Identifying and extracting data trapped in our content

Group: Josh, Greg, Marco, Danny, Cheon

Where do you store it?

Picking a specific -- Categories? But that already is a property of MediaWiki, compared to a template. The concept of category exists in the core data structures, but infobox doesn't. Categories are a semantic network. They're more like tags, not ontological.

See also is all wikitext. It's stored as raw strings. If you wanted a voice agent to read them to you one after another, that would be interesting.

Model can be a css model, series of overrides. As a local community, you can override, the way that infoboxes work for a lot of things. There's a general person infobox, then other people build on that. This is what happened when wikidata descriptions, English had a problem, we said you can override it, now it's kept locally as an unstructured magic word. I think that's a general principle that can solve this -- if there's a global default, you can do local overrides when you want it.

If you can override the semantics, it should go somewhere else, or to Wikidata. Which one is better? Don't think you can answer that in general, def not for the whole community. If we do come up with a nice way to do descriptions with Wikidata, and people say no, then we have to redo things. It's hard to say this is the place.

WPs have specific concerns about WD, and if WD wants to be the global database of all truth, then they need to address those concerns. One concern is citations, WD uses a different model. At some point, there's going to be some kind of compromise on that. If they want to store everything, they need to prioritize those issues.

How do you get stuff out of these places?

There's getting it out, and there's also getting it one time rather than getting it every time. You get the see alsos, but then that's going to change over time. Do you do it as a template?

Say you have the content translation tool -- is that a forking tool or a synching tool? Right now, we fell into a forking model, but that wasn't a deliberate decision.

It's much harder than translate, because you have to take into account the wiki, each has their own rules & way they do things. That usually means somebody started something, and

# Identifying and extracting data trapped in our content

everyone copied. 800 projects, you have to check at least 800 cases, assuming there's only one template per project.

The question should be how do we talk to all of these communities, to get a single way of doing things that are so common? There's a simple technical way, but community is hard. We struggle with -- we want these things for us for our own technical purposes, but we need to make it work for them. It needs to help them in some way. People are happy with the way they've been doing work for years.

Maybe get them to allow us to edit their edits? if you're happy with doing the template this way, then just allow me to reformat it on save, the way it's better. This is how we ended up with Michael's team, we do this structuring ourselves. We take the main pages, and his team extracts the data we need. We had to talk to individual template owners on some languages where things were especially weird. But we have to keep up with changes, it drifts over time.

Another thing that comes up, often the templates are copied. Basically, most templates are copied from english or russian, because they've gone around advocating for their templates and offering help. It's almost like a genetic process, there are two families -- the English-derived templates and Russian-derived templates.

For extraction, there's also the question of do you enter wikitext, stored in wikitext, or do you just have the semantic layer, where you can enter see alsos into a structured list? It's a big ask.

When you submit the edit -- it looks like you're trying to add structured data, can I help? It's like Clippy, but maybe it's Jimmy. :)

Funny thing is: if we just make the output look the same and we do the transition work, people would probably be fine. It's only when it visually changes that people care. :)

What needs versioning?

When things move, or there are errors. Everything we're talking about can have vandalism and problems.

For debugging purposes, it would be good to look at the parsed-out wikitext, as the template has been expanded and turned into normal storage form. But it's not deterministic. Lack of stored revisions in the parser. The main problems are the templates themselves, you can't do determ parses, bc the templates are not self-enclosed. the render is screwed up completely. so

# Identifying and extracting data trapped in our content

you can't have structured parsing. some exceptions where people refuse to correct or just don't know it's a problem. structured templates would be much easier. wikitext 2.0!

maybe that's something for this whole discussion, rather than think about structuring individual things, make the template system first, and all you're doing is editing wikitext for the template. Most of these items are in some kind of template anyway. except for see also. portals are all templates.

do we have project group for category structure on WP? not really. they're so complex and used in so many different ways, we haven't tried to get into it. no group on english, a category project? There are category minders, who argue abou the structure -- but mostly in specific topic areas. categories are also used for things that aren't content, marking things for articles for deletion. it's so general purpose that we don't want to change it much, anything we do would break another use case.