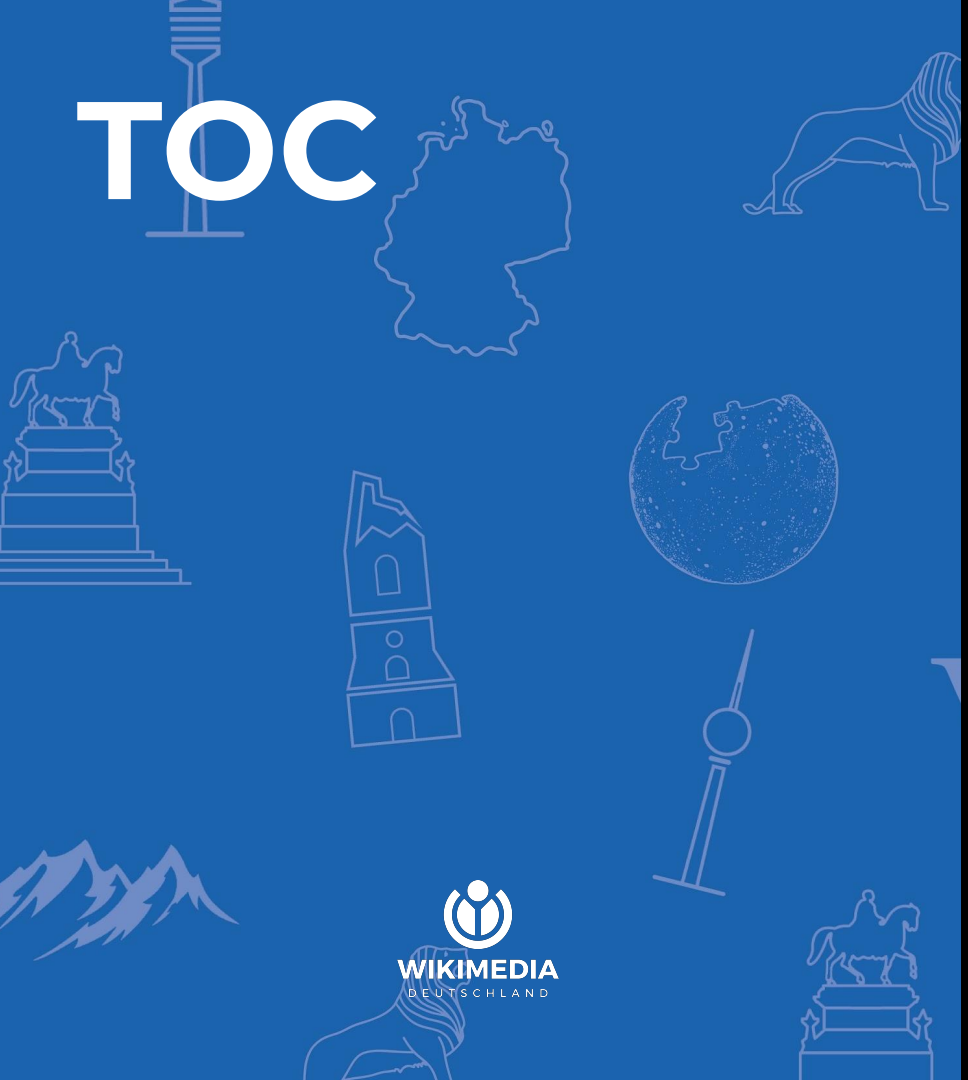
Wikidata, behind the curtain

Amir Sarabadani
amir.sarabadani@wikimedia.de

WIKIMEDIA
DEUTSCHLAND

TOC



- Concepts
- Code structure
- Representations/API
- Secondary data storages
- Entity Usage
- Front-end stack
- Miscellaneous parts
- Future changes

Concepts

WIKIMEDIA
DEUTSCHLAND

Concepts

- **Wikibase:**
  - Wikibase Repo: Wikidata, Wikimedia Commons
  - Wikibase Client: English Wikipedia, German Wikisource, Wikidata
  - Federation
- **Entities:**
  - Item
  - Property
- **Data model:**
  - Terms: labels, descriptions, aliases
  - Statements
  - Sitelinks
- **Other Entities:**
  - Lexeme
    i. Basically dictionary entries
    ii. Has a different data model
  - MediaInfo
    i. One per file in commons, making Structured Data on Commons possible
    ii. Has a different data model

Alan Turing (Q7251)

British mathematician, logician, cryptanalyst, and computer scientist

Alan M. Turing | Alan Mathieson Turing | Turing | Alan Mathison Turing

Terms

aliases

In more languages Configure

Language	Label	Description	Also known as
English	Alan Turing	British mathematician, logician, cryptanalyst, and computer scientist	Alan M. Turing Alan Mathieson Turing Turing Alan Mathison Turing
Persian	آلن تورینگ	دانشمند کامپیوتر، رمزشکن، منطق‌دان و ریاضی‌دان بریتانیایی	
Polish	Alan Turing	angielski matematyk	
German	Alan Turing	britischer Logiker, Mathematiker und Kryptoanalytiker	Alan Mathison Turing Alan M. Turing

All entered languages

Statements

instance of	human 3 references

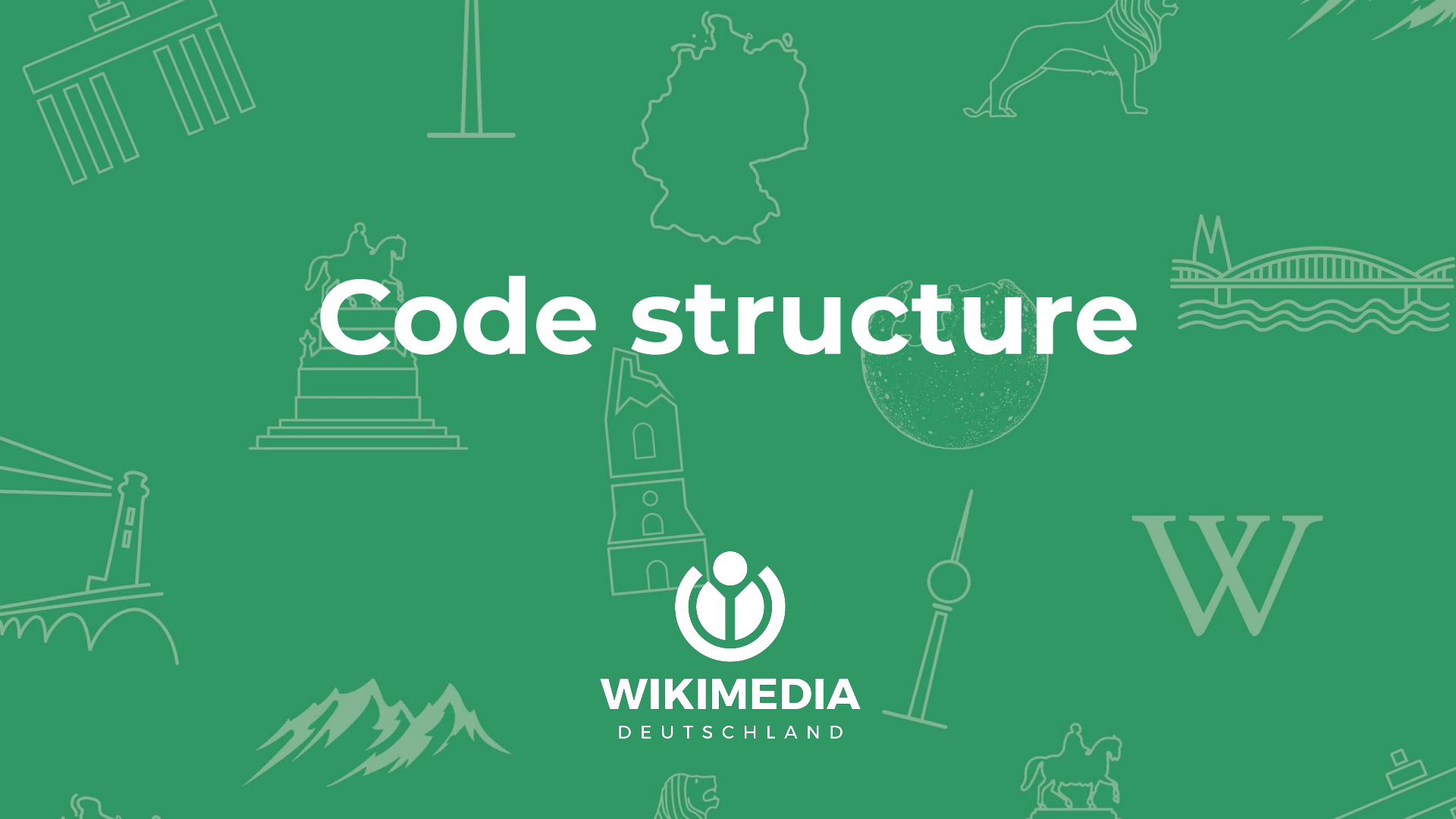
Statements

Wikipedia (141 entries)

- ady Тьюринг Алан
- af Alan Turing
- als Alan Turing
- an Alan Turing
- ar آلان تورنغ
- arz الان تورينج
- ast Alan Turing
- as এলান ট্যুৰিং
- azb آلن تورینق
- az Alan Türinq
- bat_smg Alans Tiorėngs
- ba Алан Тьюринг
- bcl Alan Turing
- be_x_old Элан Т'юрынг
- be Алан Мэтысан Цьюрынг
- bg Алън Тюринг
- bh एलन ट्यूरिंग
- bn অ্যালান টুরিং
- br Alan Turing
- bs Alan Turing

Sitelinks

Code structure



Example: Wikimedia Commons
(A repo wiki)

Wikibase				
Extension	**Version**	**License**	**Description**	**Authors**
WikibaseCirrusSearch	0.1.0 (788d536) 21:28, 4 September 2019	GPL-2.0-or-later	Wikibase extension for using CirrusSearch functionality	The Wikidata team and Stas Malyshev
WikibaseClient	– (2c7bfb6) 16:04, 1 October 2019	GPL-2.0-or-later	Client for the Wikibase extension	The Wikidata team and others
WikibaseLexeme	– (47a8eb7) 11:08, 1 October 2019	GPL-2.0-or-later	Handle Lexeme entity type	The Wikidata team
WikibaseLib	– (2c7bfb6) 16:04, 1 October 2019	GPL-2.0-or-later	Holds common functionality for the Wikibase and Wikibase Client extensions	The Wikidata team and others
WikibaseMediaInfo	0.1.0 (6555e13) 07:27, 1 October 2019	GPL-2.0-or-later	Wikibase extension to manage structured metadata of media files	The Wikidata team
WikibaseRepository	– (2c7bfb6) 16:04, 1 October 2019	GPL-2.0-or-later	Structured data repository	The Wikidata team and others
WikibaseView	– (2c7bfb6) 16:04, 1 October 2019	GPL-2.0-or-later	View component for the Wikibase Repository	The Wikidata team and others
WikimediaBadges	1.0.0 (a306315) 11:35, 23 August 2019	GPL-2.0-or-later	Default themes to display badges on Wikimedia projects	Bene* and Marius Hoch

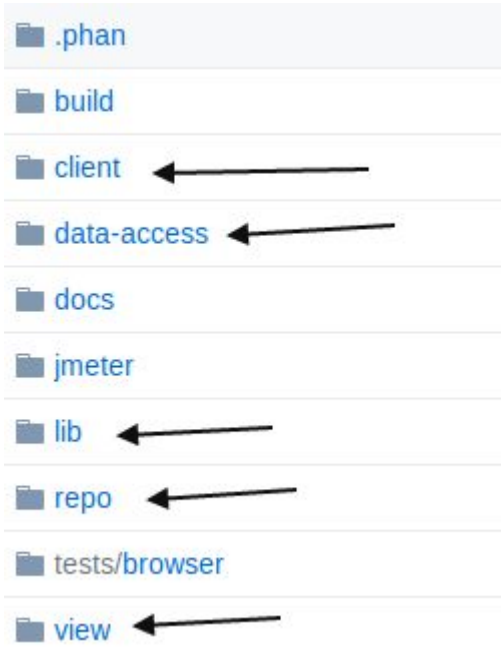
WIKIMEDIA
DEUTSCHLAND

Example: English Wikipedia
(A client wiki)

Wikibase				
Extension	**Version**	**License**	**Description**	**Authors**
WikibaseClient	– (2c7bfb6) 16:04, 1 October 2019	GPL-2.0-or-later	Client for the Wikibase extension	The Wikidata team and others
WikibaseLexeme	– (47a8eb7) 11:08, 1 October 2019	GPL-2.0-or-later	Handle Lexeme entity type	The Wikidata team
WikibaseLib	– (2c7bfb6) 16:04, 1 October 2019	GPL-2.0-or-later	Holds common functionality for the Wikibase and Wikibase Client extensions	The Wikidata team and others
WikimediaBadges	1.0.0 (a306315) 11:35, 23 August 2019	GPL-2.0-or-later	Default themes to display badges on Wikimedia projects	Bene* and Marius Hoch

WIKIMEDIA
DEUTSCHLAND

Directories



“data-access” is part of “lib”
“view” is part of “repo”

WIKIMEDIA
DEUTSCHLAND

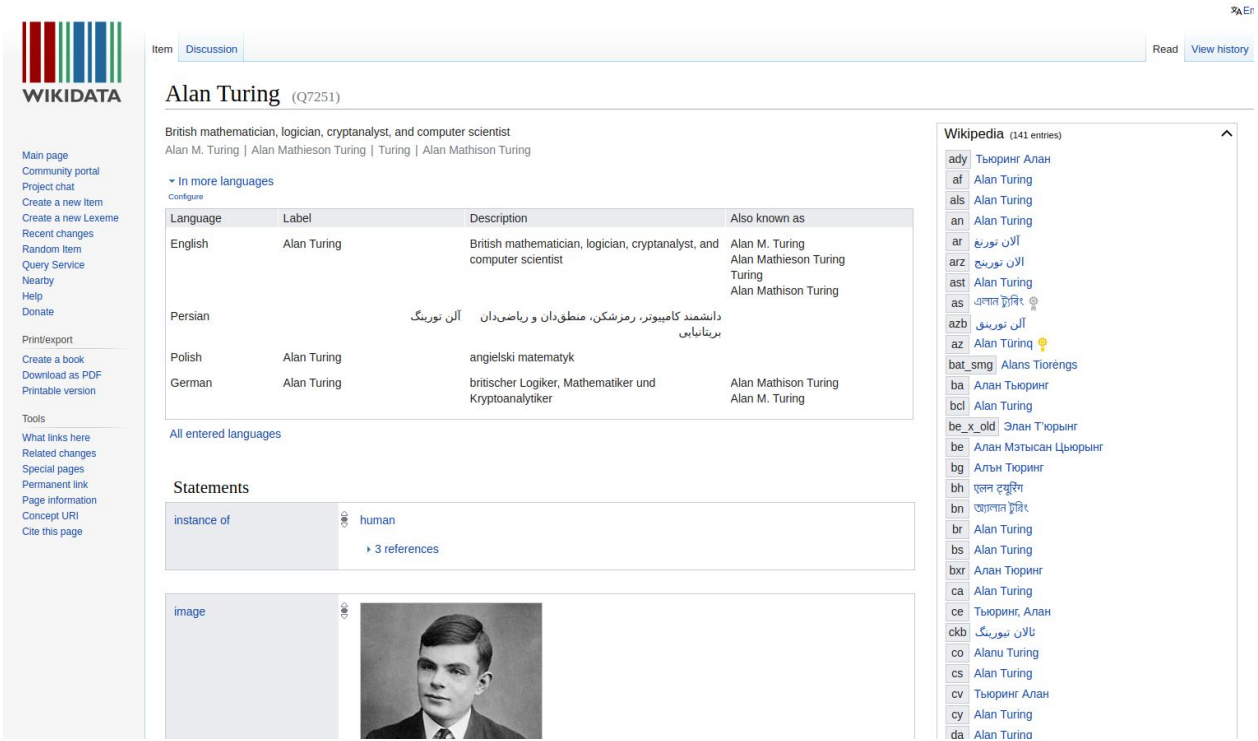
Representations/APIs

One source of truth
(Internal)

```
ladsgroup@mwmaint1002:~$ mwscript getText.php wikidatawiki 'Q7251'
{"type":"item","id":"Q7251","labels":{"fr":{"language":"fr","value":"Alan Turing
"},"en":{"language":"en","value":"Alan Turing"},"pl":{"language":"pl","value":"A
lan Turing"},"it":{"language":"it","value":"Alan Turing"},"ilo":{"language":"ilo
","value":"Alan Turing"},"jbo":{"language":"jbo","value":".alan turin"},"be-tara
sk":{"language":"be-tarask","value":"\u042d\u043b\u0430\u043d \u0422\u2019\u044e
\u0440\u044b\u043d\u0433"},"en-ca":{"language":"en-ca","value":"Alan Turing"},"e
n-gb":{"language":"en-gb","value":"Alan Turing"},"af":{"language":"af","value":"
Alan Turing"},"an":{"language":"an","value":"Alan Turing"},"ar":{"language":"ar"
,"value":"\u0622\u0644\u0627\u0646 \u062a\u0648\u0631\u0646\u063a"},"arz":{"lang
uage":"arz","value":"\u0627\u0644\u0627\u0646 \u062a\u0648\u0631\u064a\u0646\u06
2c"},"as":{"language":"as","value":"\u098f\u09b2\u09be\u09a8 \u099f\u09cd\u09af\
u09c1\u09f0\u09bf\u0982"},"ast":{"language":"ast","value":"Alan Turing"},"az":{"
language":"az","value":"Alan T\u00fcrinq"},"be":{"language":"be","value":"\u0410
\u043b\u0430\u043d \u041c\u0430\u0442\u044b\u0441\u0430\u043d \u0426\u044c\u044e
\u0440\u044b\u043d\u0433"},"bg":{"language":"bg","value":"\u0410\u043b\u044a\u04
3d \u0422\u044e\u0440\u0438\u043d\u0433"},"bn":{"language":"bn","value":"\u0985\
u09cd\u09af\u09be\u09b2\u09be\u09a8 \u099f\u09c1\u09b0\u09bf\u0982"},"br":{"lang
```

HTML

/wiki/Special:EntityData/Q7251
/wiki/Q7251



JSON API

/wiki/Special:EntityData/Q7251.json
/w/api.php?action=wbgetentities&id=Q7251


```
entities:
  Q7251:
    pageid:          8455
    ns:              0
    title:           "Q7251"
    lastrevid:       1023644940
    modified:        "2019-10-01T04:00:08Z"
    type:            "item"
    id:              "Q7251"
  labels:
    fr:
      language:      "fr"
      value:         "Alan Turing"
    en:
      language:      "en"
      value:         "Alan Turing"
    pl:
      language:      "pl"
      value:         "Alan Turing"
    it:
      language:      "it"
      value:         "Alan Turing"
```


WIKIMEDIA
DEUTSCHLAND

RDF (ttl, jsonld)

/wiki/Special:EntityData/Q7251.ttl

```
@prefix pr: <http://www.wikidata.org/prop/reference/> .
@prefix prv: <http://www.wikidata.org/prop/reference/value/> .
@prefix prn: <http://www.wikidata.org/prop/reference/value-normalized/> .
@prefix wdno: <http://www.wikidata.org/prop/novalue/> .

data:Q7251 a schema:Dataset ;
    schema:about wd:Q7251 ;
    cc:license <http://creativecommons.org/publicdomain/zero/1.0/> ;
    schema:softwareVersion "1.0.0" ;
    schema:version "1023644940"^^xsd:integer ;
    schema:dateModified "2019-10-01T04:00:08Z"^^xsd:dateTime ;
    wikibase:statements "141"^^xsd:integer ;
    wikibase:identifiers "64"^^xsd:integer ;
    wikibase:sitelinks "159"^^xsd:integer .

wd:Q7251 a wikibase:Item .

<https://commons.wikimedia.org/wiki/Alan_Turing> a schema:Article ;
    schema:about wd:Q7251 ;
    schema:inLanguage "en" ;
    schema:isPartOf <https://commons.wikimedia.org/> ;
    schema:name "Alan Turing"@en .

<https://commons.wikimedia.org/> wikibase:wikiGroup "commons" .

<https://fr.wikiquote.org/wiki/Alan_Turing> a schema:Article ;
```

/wiki/Special:EntityData/Q7251.jsonld


```
@graph:
  0:
    @id:              "data:Q7251"
    @type:            "schema:Dataset"
    about:            "wd:Q7251"
    license:          "http://creativecommons.org/publicdomain/zero/1.0/"
    softwareVersion:  "1.0.0"
    version:          1023644940
    dateModified:     "2019-10-01T04:00:08Z"
    statements:       141
    identifiers:      64
    sitelinks:        159
  1:
    @id:              "wd:Q7251"
    @type:            "wikibase:Item"
  2:
    @id:              "https://commons.wikimedia.org/wiki/Alan_Turing"
    @type:            "schema:Article"
    about:            "wd:Q7251"
    inLanguage:       "en"
    isPartOf:         "https://commons.wikimedia.org/"
    name:
      @language:      "en"
      @value:         "Alan Turing"
  3:
    @id:              "https://commons.wikimedia.org/"
    wikiGroup:        "commons"
```

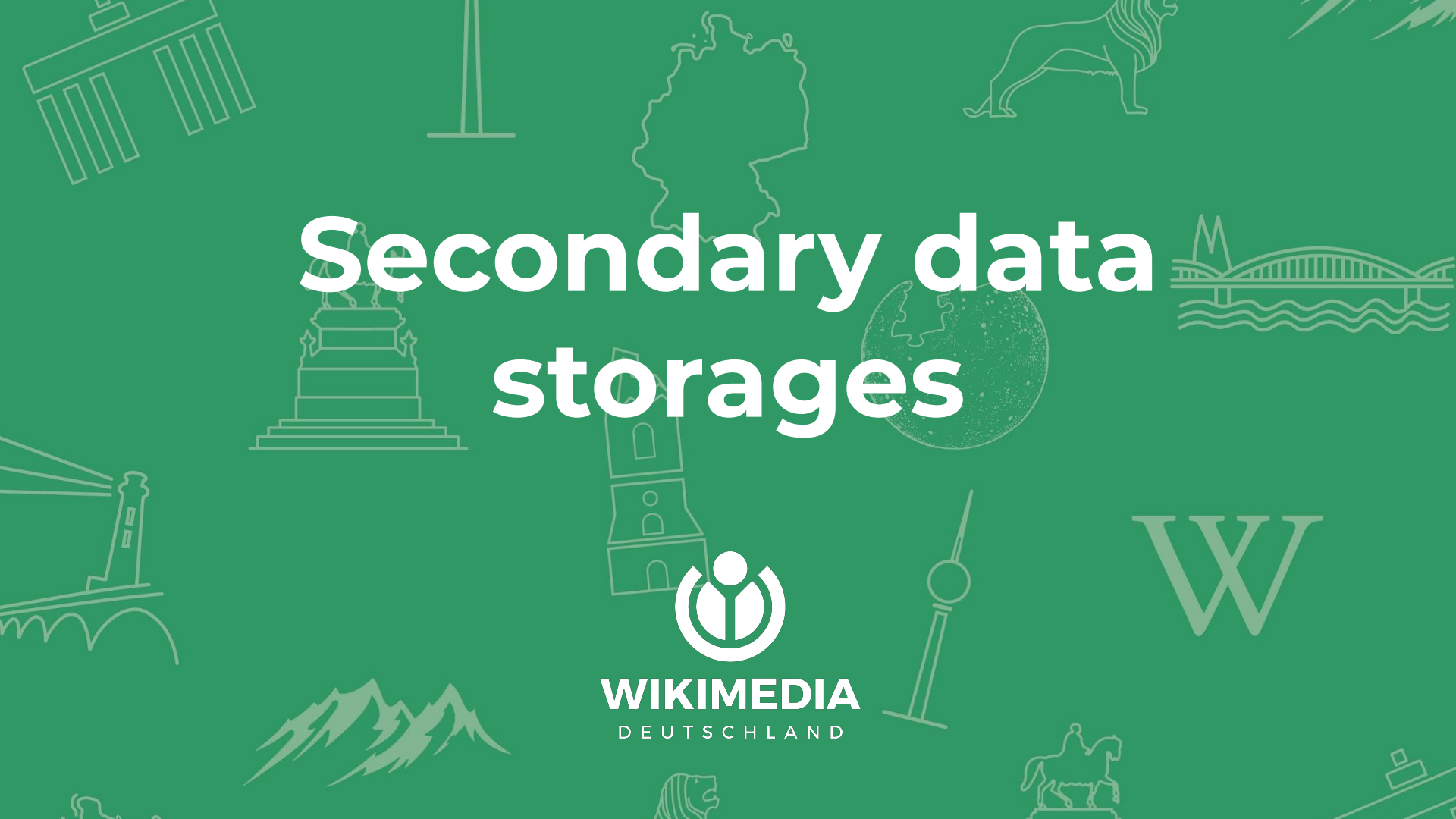

WIKIMEDIA DEUTSCHLAND

PHP API

```
ladsgroup@mwmaint1002:~$ mwscript eval.php --wiki=wikidatawiki
> $entityLookup = \Wikibase\Repo\WikibaseRepo::getDefaultInstance()->getEntityLookup();

> $item = $entityLookup->getEntity( new \Wikibase\DataModel\Entity\ItemId( 'Q7251') );

> var_dump( $item->getLabels()->getByLanguage( 'en' ) );
object(Wikibase\DataModel\Term\Term)#811 (2) {
  ["languageCode":"Wikibase\DataModel\Term\Term":private]=>
  string(2) "en"
  ["text":"Wikibase\DataModel\Term\Term":private]=>
  string(11) "Alan Turing"
}

>
```

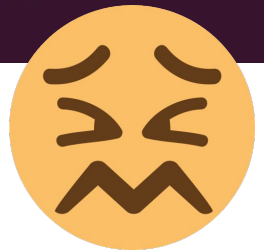
WIKIMEDIA
DEUTSCHLAND

Secondary data storages



Term store
(+ its caches)

```
wikiadmin@10.64.48.151(wikidatawiki)> select * from wb_terms where term_full_entity_id = 'Q7251' limit 5;
+-------------+---------------+--------------------+-----------------+---------------+------------+-------------------------------------------------------------------------+-----------------------------
-------------------------------------------+------------+
| term_row_id | term_entity_id | term_full_entity_id | term_entity_type | term_language | term_type   | term_text                                                                | term_search_key
                                           | term_weight |
+-------------+---------------+--------------------+-----------------+---------------+------------+-------------------------------------------------------------------------+-----------------------------
-------------------------------------------+------------+
|  1243775063 |           7251 | Q7251              | item            | ilo           | description | Britaniko a matematiko, lohiko, kriptoanalista, ken sientista ti kompiuter | britaniko a matematiko, lohiko
, kriptoanalista, ken sientista ti kompiuter |      0.152 |
|  1243775069 |           7251 | Q7251              | item            | pl            | description | angielski matematyk                                                      | angielski matematyk
                                           |      0.152 |
|  1243775070 |           7251 | Q7251              | item            | sk            | description | britský matematik, logik, kryptograf a vojnový hrdina                    | britský matematik, logik, kryp
tograf a vojnový hrdina                    |      0.152 |
|  1243775071 |           7251 | Q7251              | item            | as            | description | Computer scientist, mathematician, and cryptographer                     | computer scientist, mathematic
ian, and cryptographer                     |      0.152 |
|  1243775072 |           7251 | Q7251              | item            | sv            | description | brittisk datavetare, matematiker och ingenjör                            | brittisk datavetare, matematik
er och ingenjör                            |      0.152 |
+-------------+---------------+--------------------+-----------------+---------------+------------+-------------------------------------------------------------------------+-----------------------------
-------------------------------------------+------------+
5 rows in set (0.00 sec)

wikiadmin@10.64.48.151(wikidatawiki)>
```



Credit: Emoji one,
CC-BY-SA 4.0

WIKIMEDIA
DEUTSCHLAND

ParserCache

- Stores the HTML representation
- Plugs into and the same as MediaWiki ParserCache
- Includes placeholders, cached values for information substituted per user preferences at read time
- Entry point for server-side rendering of VueJS micro front ends.
- Expires in a 30 days or gets invalidated by edit
- Fragmented by language (different than most WMF Wikis)

WIKIMEDIA
DEUTSCHLAND

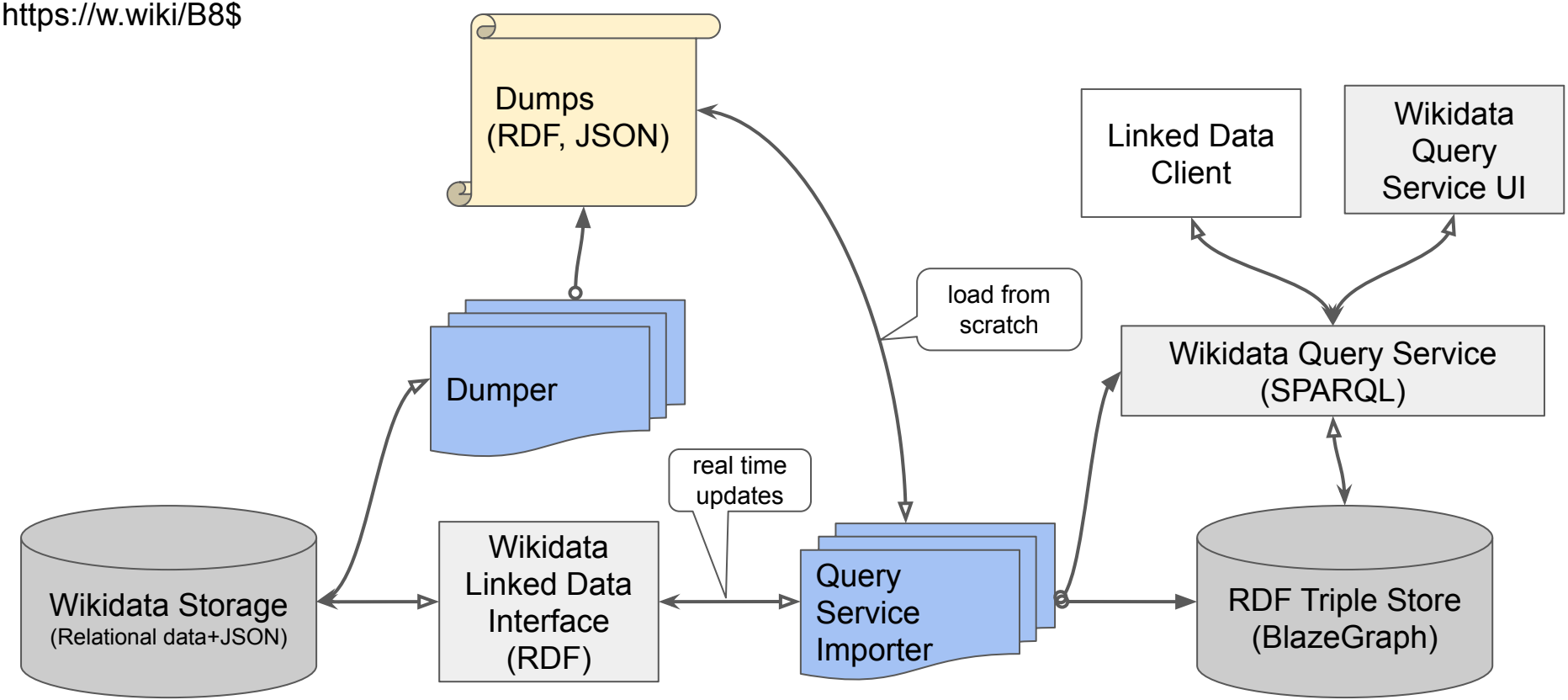
Blazegraph

- Stores an optimized version of RDF (turtle) representation
- That's what empowers Wikidata query service (WDQS, https://query.wikidata.org)
- A graph database made using RDF (turtle) representation of entities
- Completely out of appservers and wikidata itself.

WIKIMEDIA
DEUTSCHLAND

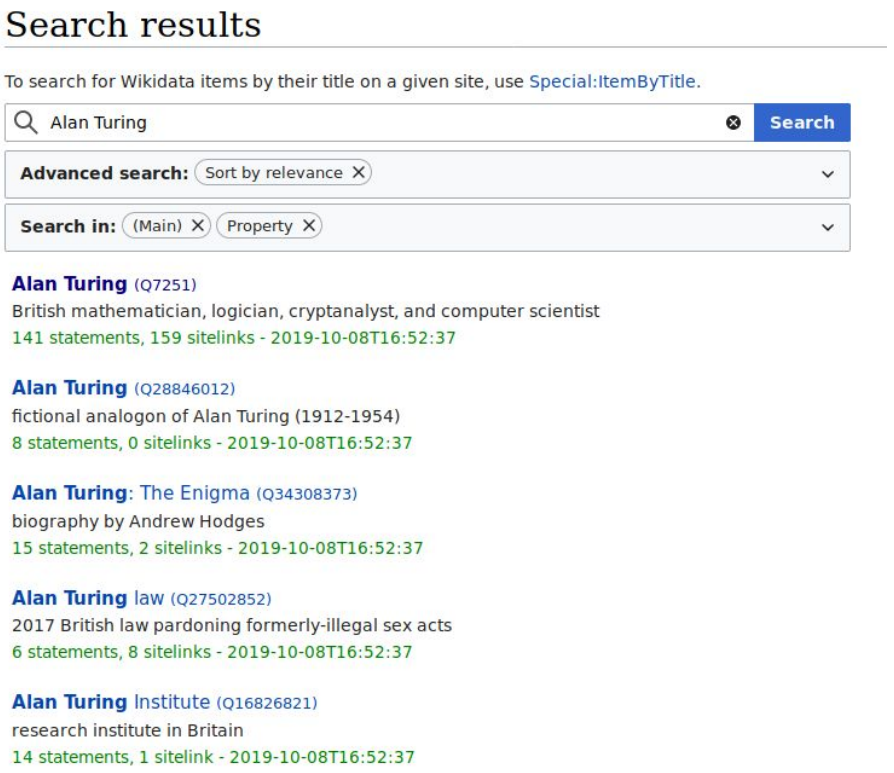
WDQS infrastructure

https://w.wiki/B8$



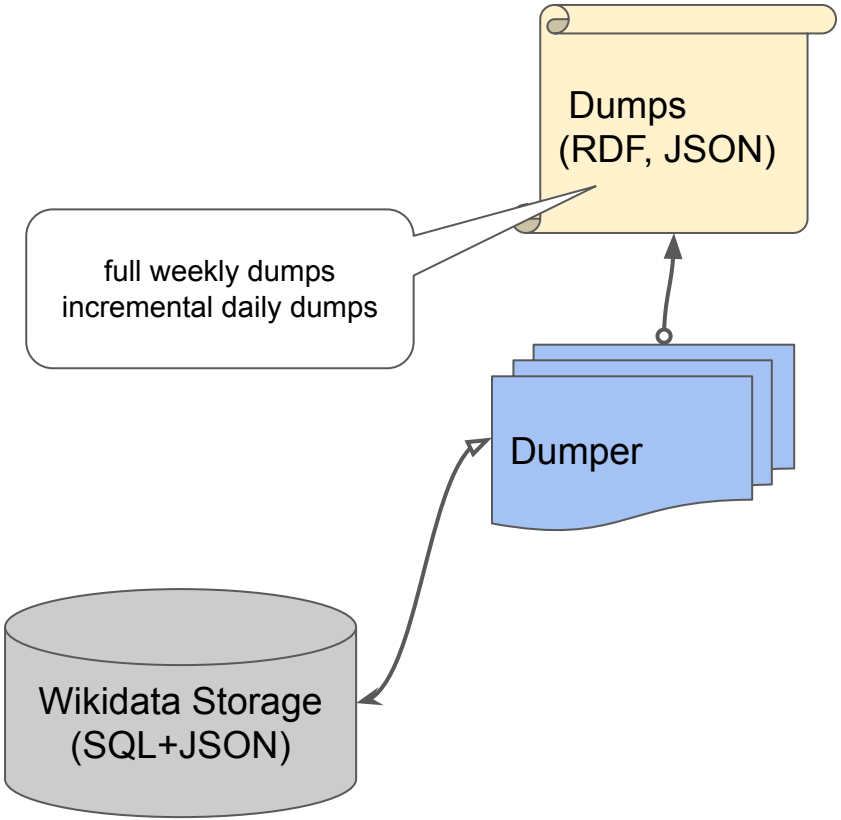
ElasticSearch

- Handled through WikibaseCirrusSearch extension
- It's optional, third parties can use the term store instead.
- Exposed via Special:Search (duh) and wbsearchentities API module



WIKIMEDIA
DEUTSCHLAND

Dumps


Dumps
(RDF, JSON)
full weekly dumps
incremental daily dumps
Dumper
Wikidata Storage
(SQL+JSON)

Entity usage

On Repo

- Rendering another item
- Special pages

WIKIMEDIA
DEUTSCHLAND

On Client

- Sitelink usage
- Data usage
  - Parser functions
  - Lua functions (https://w.wiki/9Va)

WIKIMEDIA DEUTSCHLAND

https://w.wiki/9VZ

Tracking entity usage

Client-side:

```
MariaDB [euwiki_p]> select * from wbc_entity_usage where eu_entity_id = 'Q7251';
+------------+--------------+-----------+------------+
| eu_row_id  | eu_entity_id | eu_aspect | eu_page_id |
+------------+--------------+-----------+------------+
|    5792173 | Q7251        | C         |      24076 |
|   29936634 | Q7251        | L.eu      |      24076 |
|    2961057 | Q7251        | L.eu      |     284324 |
|   26509057 | Q7251        | L.eu      |     580923 |
|   18681943 | Q7251        | L.eu      |     807797 |
|     636108 | Q7251        | O         |      24076 |
|     636111 | Q7251        | S         |      24076 |
|     636109 | Q7251        | T         |      24076 |
|    2961059 | Q7251        | T         |     284324 |
|   26509058 | Q7251        | T         |     580923 |
+------------+--------------+-----------+------------+
10 rows in set (0.01 sec)

MariaDB [euwiki_p]>
```

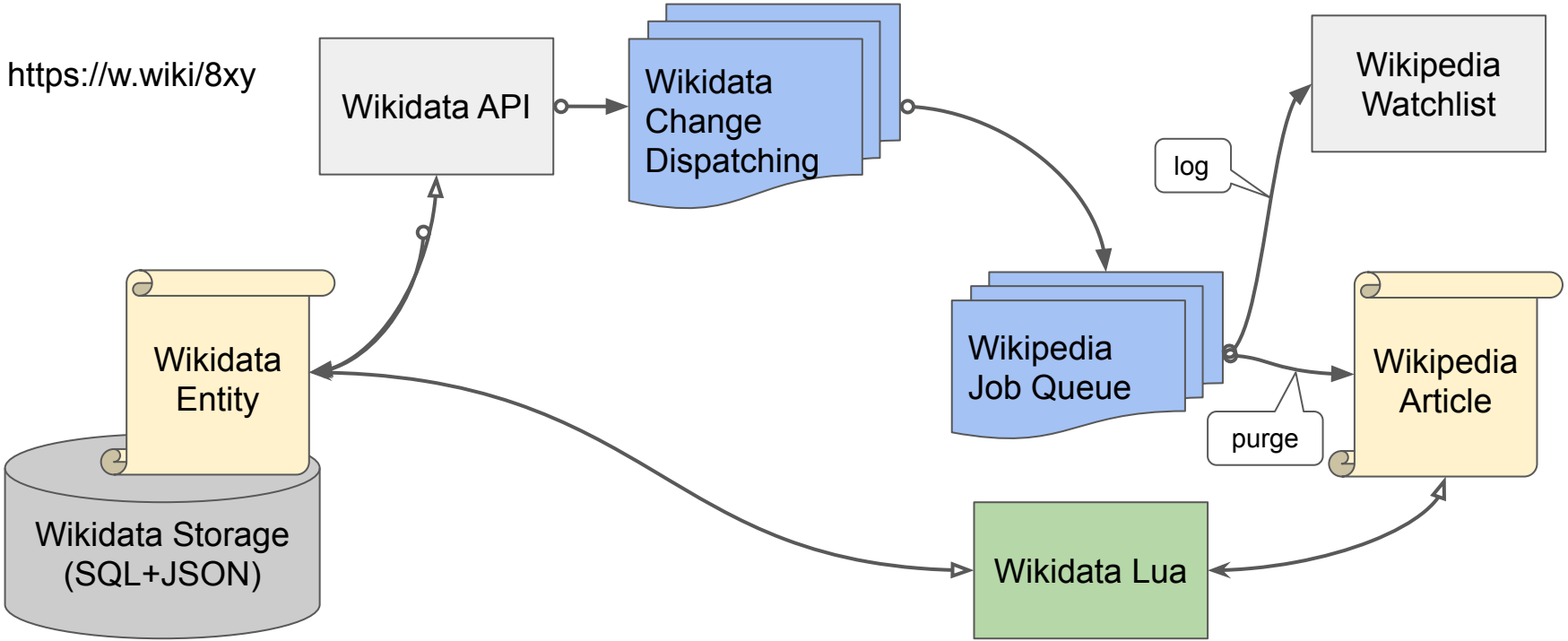
Repo-side:

```
MariaDB [wikidatawiki_p]> select * from wb_changes_subscription where cs_entity_id = 'Q7251';
+-----------+--------------+------------------+
| cs_row_id | cs_entity_id | cs_subscriber_id |
+-----------+--------------+------------------+
| 103425677 | Q7251        | adywiki          |
|  57014298 | Q7251        | afwiki           |
|  57041654 | Q7251        | alswiki          |
|  57095558 | Q7251        | anwiki           |
|  49287167 | Q7251        | arwiki           |
|  57122747 | Q7251        | arzwiki          |
|  57145349 | Q7251        | astwiki          |
|  57154877 | Q7251        | aswiki           |
| 113578216 | Q7251        | azbwiki          |
|  76413512 | Q7251        | azwiki           |
|  57318864 | Q7251        | bat_smgwiki      |
|  57367856 | Q7251        | bawiki           |
| 117022292 | Q7251        | bclwiki          |
|  57455739 | Q7251        | be_x_oldwiki     |
| 106847561 | Q7251        | bewiki           |
| 106849366 | Q7251        | bewikiquote      |
|  72176889 | Q7251        | bgwiki           |
| 107711230 | Q7251        | bhwiki           |
|  57669624 | Q7251        | bnwiki           |
|  57771676 | Q7251        | brwiki           |
|  57863678 | Q7251        | bswiki           |
| 102052821 | Q7251        | bxrwiki          |
```

WIKIMEDIA DEUTSCHLAND

Change dispatching

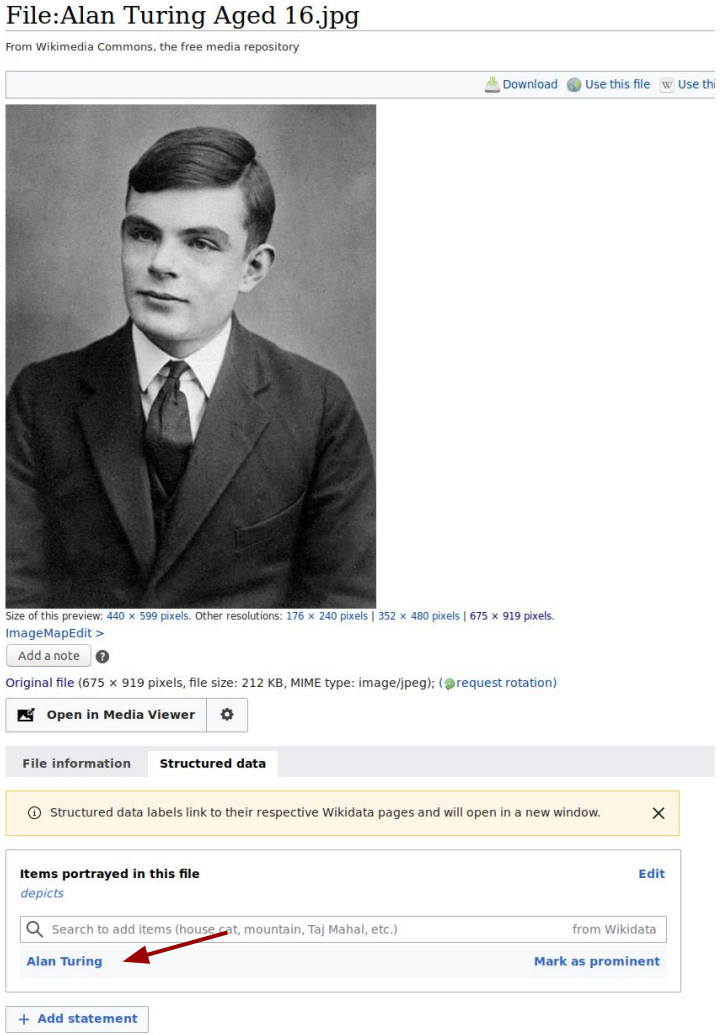
https://w.wiki/8xy



On another repo


```
entities:
  M22828488:
    pageid:                 22828488
    ns:                     6
    title:                  "File:Alan Turing Aged 16.jpg"
    lastrevid:              357356024
    modified:               "2019-07-07T12:00:20Z"
    type:                   "mediainfo"
    id:                     "M22828488"
    labels:                 {}
    descriptions:           {}
    statements:
      P180:
        0:
          mainsnak:
            snaktype:       "value"
            property:       "P180"
            datavalue:
              value:
                entity-type:    "item"
                numeric-id:     7251
                id:             "Q7251"
              type:         "wikibase-entityid"
          type:             "statement"
          id:               "M22828488$1c87d708-4dca-385a-d80e-7bbbff7f2615"
          rank:             "normal"
```


WIKIMEDIA DEUTSCHLAND

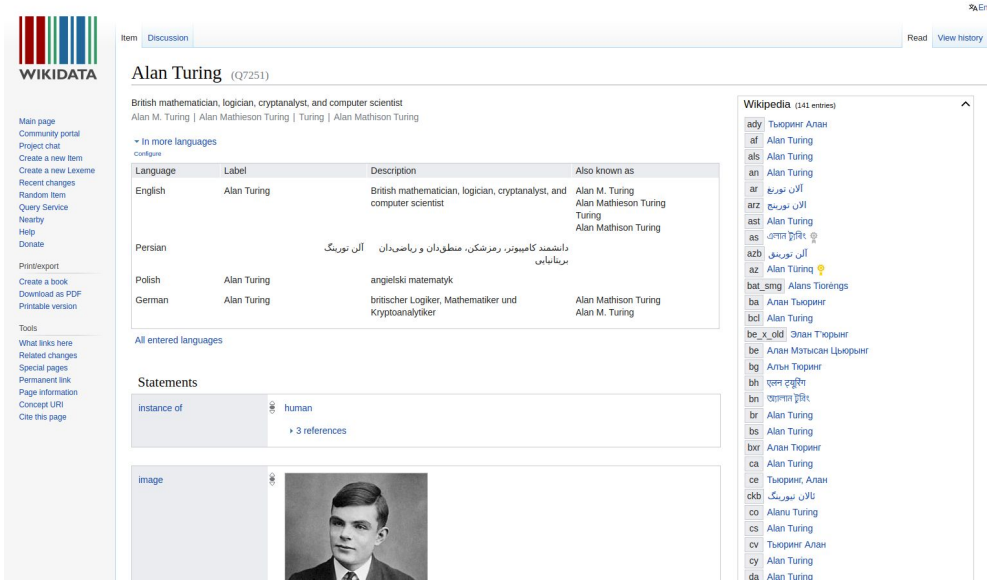


Front-end stack

WIKIMEDIA
DEUTSCHLAND

Entity pageview frontend

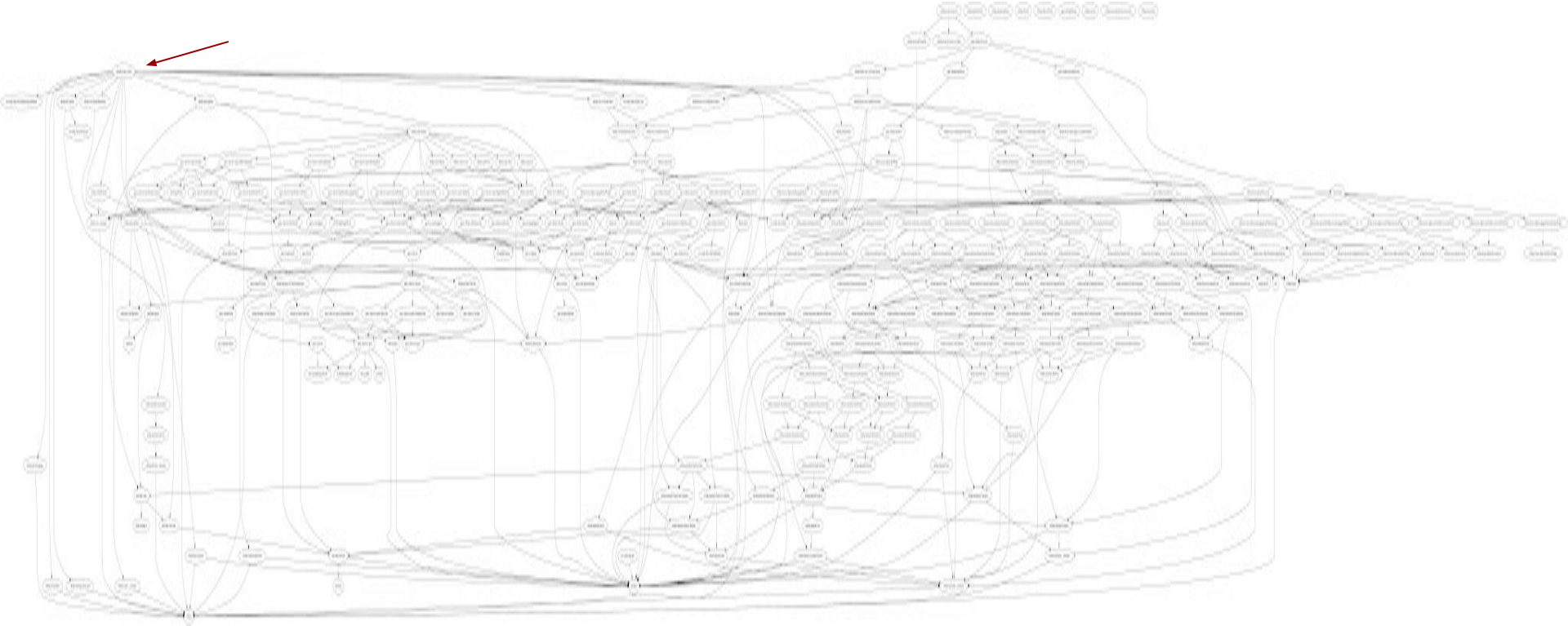
- jquery ui widgets
- Hooked into mediawiki ResourceLoader:
  - Start point: “wikibase.ui.entityViewInit” module



WIKIMEDIA DEUTSCHLAND

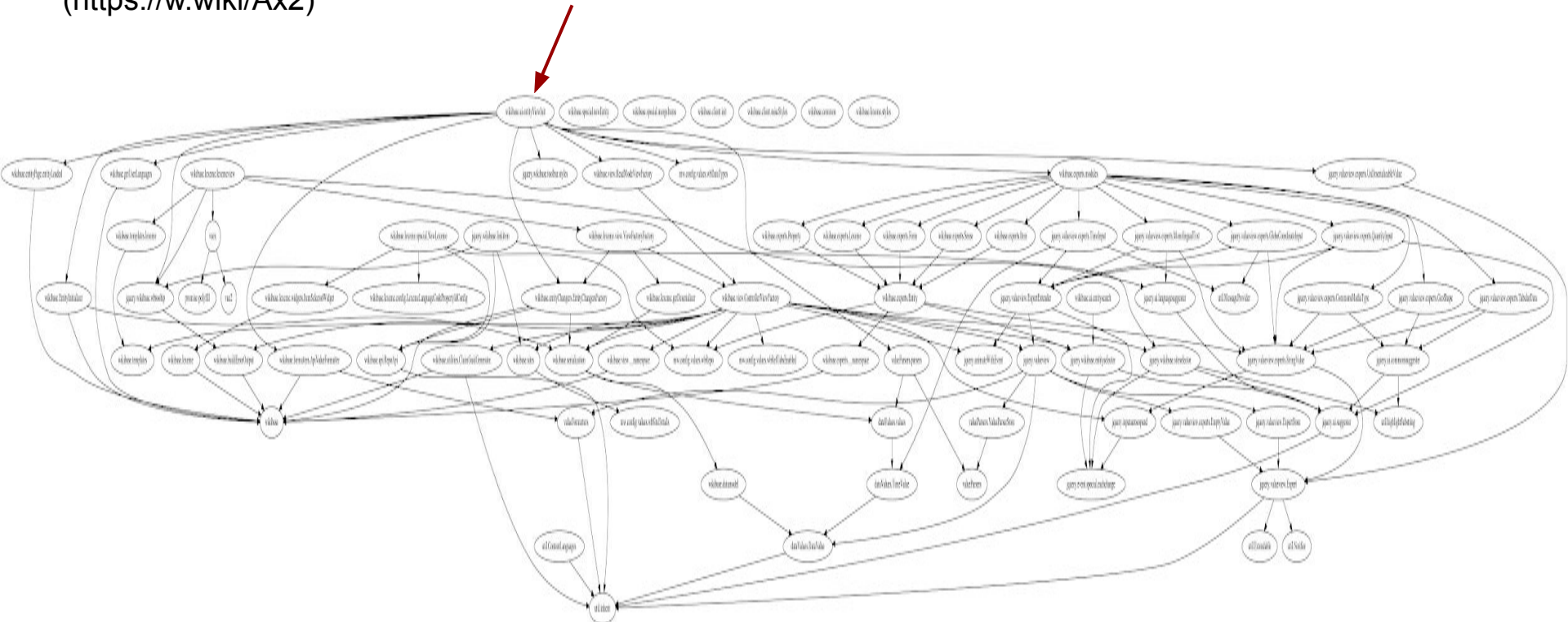
Legacy

(https://w.wiki/9VN)



Refactoring RL modules

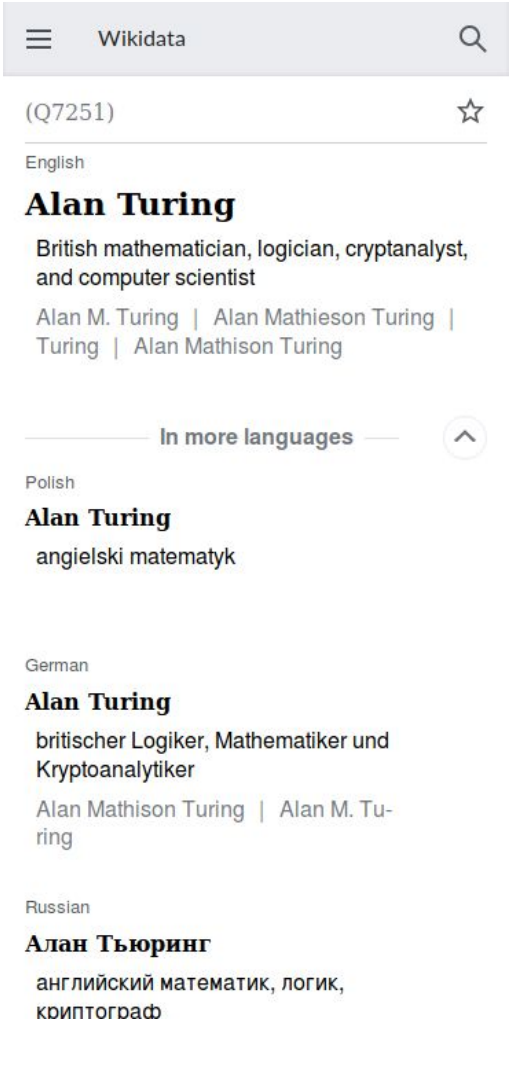
(https://w.wiki/Ax2)



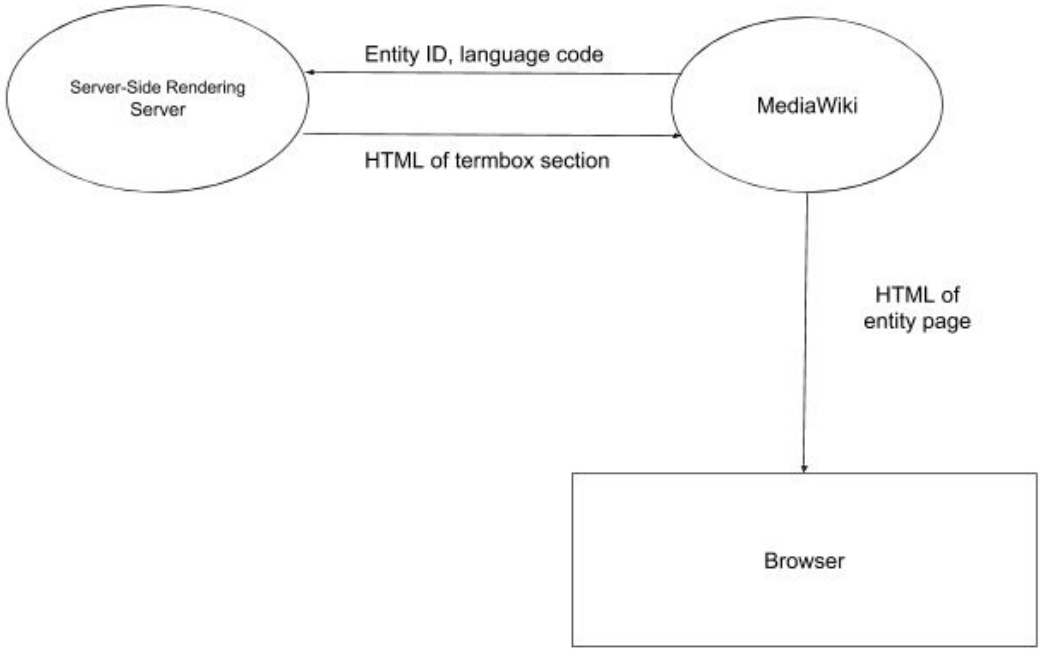
Modernizing frontend

- Server Side Rendering
- VueJS + TypeScript

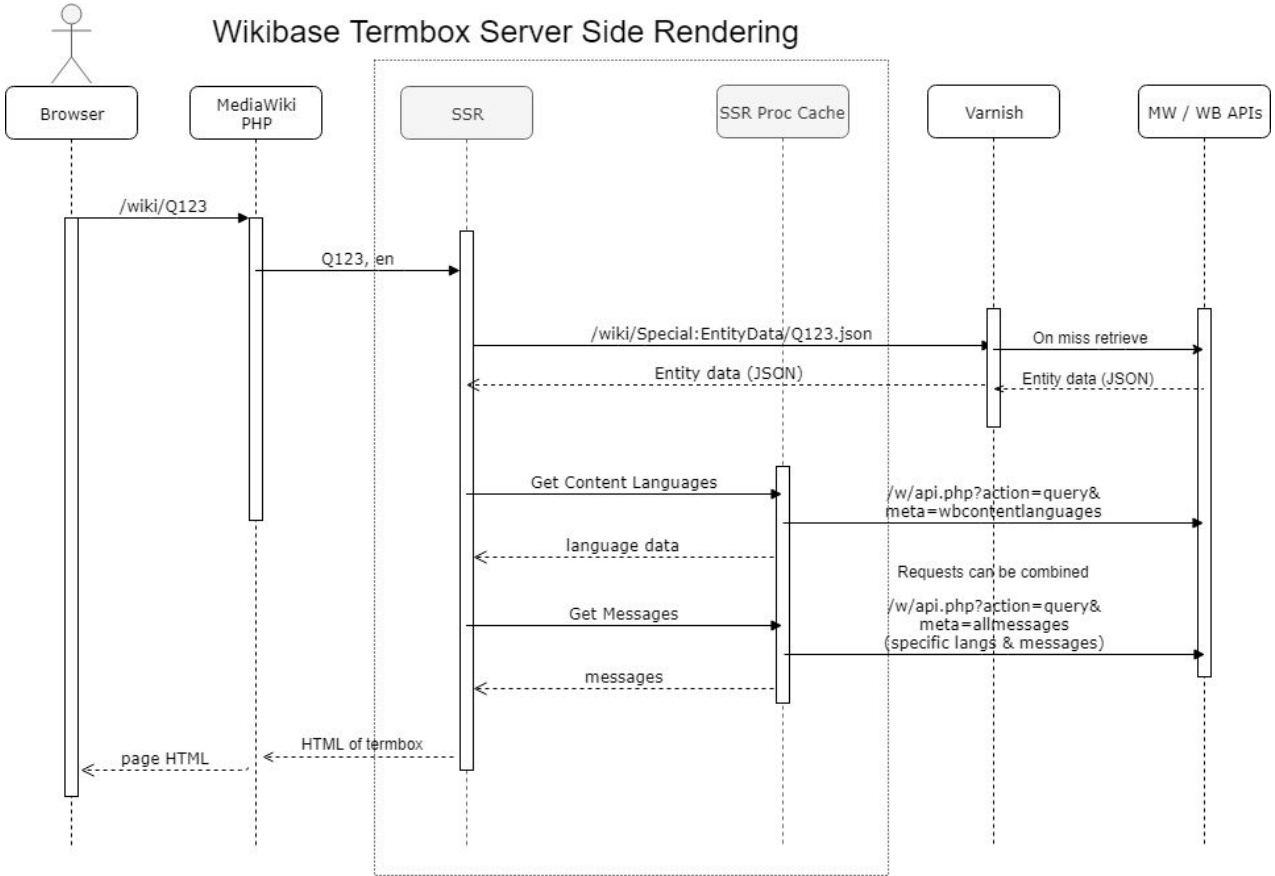
WIKIMEDIA DEUTSCHLAND



SSR, overview



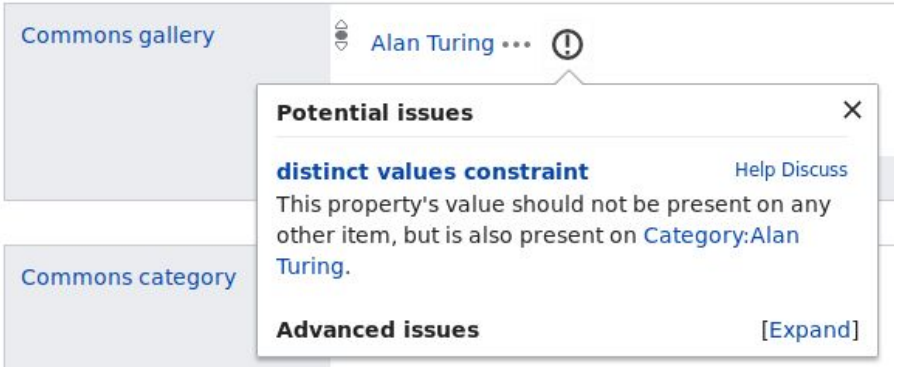
SSR, details


Wikibase Termbox Server Side Rendering
Browser
MediaWiki PHP
SSR
SSR Proc Cache
Varnish
MW / WB APIs
/wiki/Q123
Q123, en
/wiki/Special:EntityData/Q123.json
On miss retrieve
Entity data (JSON)
Entity data (JSON)
Get Content Languages
/w/api.php?action=query&
meta=wbcontentlanguages
language data
Requests can be combined
Get Messages
/w/api.php?action=query&
meta=allmessages
(specific langs & messages)
messages
HTML of termbox
page HTML

Miscellaneous parts

WIKIMEDIA
DEUTSCHLAND

Even more things to support Wikidata



- WikibaseQualityConstraints
- WMDE analytics refinery
- PropertySuggester
- ArticlePlaceholder (for client wikis)
- Bot frameworks:
  - QuickStatements
  - Pywikibot
  - WikidataIntegrator
  - ...

WIKIMEDIA
DEUTSCHLAND

Future changes

WIKIMEDIA
DEUTSCHLAND

Some changes

maintenance:

- Term store migration
- Wikidata query service backend scaling
- Frontend improvements

feature:

- Client editing (wikidata bridge)

WIKIMEDIA
DEUTSCHLAND

Docs

- Docs directory in the git repo
- Wikidata page in wikitech
- Wikibase page in mediawiki.org

WIKIMEDIA
DEUTSCHLAND

Thank you!

Amir Sarabadani
amir.sarabadani@wikimedia.de

WIKIMEDIA
DEUTSCHLAND