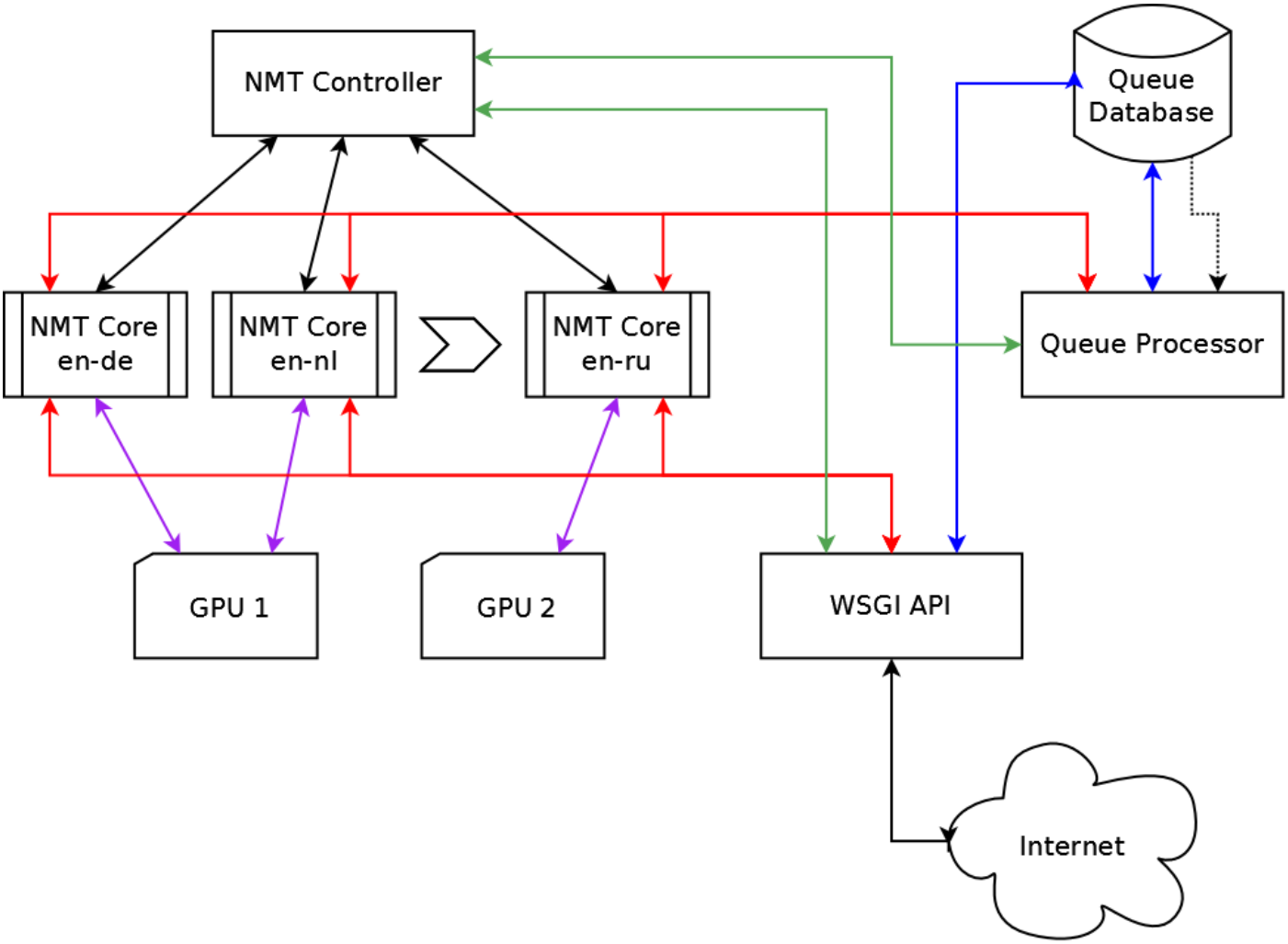
Web Service API

Web Service API v1.02

Web Service is implemented as a Python Web Server Gateway Interface (WSGI) and a set of processes controlling data flow:



WSGI application is accepting API calls, which are then relayed:

- via NMT Controller to NMT Core engines for direct translation or
- to database for queued translation

NMT Controller is responsible for starting new NMT Core engines with proper language pairs on the next free GPU or reusing the existing ones, returning the connection parameters for them. There can be several NMT Core engines (each with different language pair) running in parallel on different GPU cards as long as there is sufficient memory left on GPU cards. With the current configuration of 2 NVIDIA GeForce GTX 1080Ti with 11GB memory each there can be 8-9 concurrent NMT Core engines running in parallel.

Queue Processor listens for notifications from database (or after timeout) and process new requests from the database queue via NMT Controller and NMT Core engines in the same way as WSGI application for direct translation. After the request has been processed, the callback (if specified in the API call) is executed. If there was no callback, one can get the result of translation via another API call.

In the case of direct translation, the translation is returned as text (or in original format) and in all other cases the relevant data is returned as JSON data structure. If there has been an error processing request or some of the data hasn't been found, an appropriate HTTP 500 or 404 error is returned with more detailed explanation in the body of the returned HTML.

Base URL for TraMOOC Web service API is http://matterhorn.ijs.si/trans
On top of that API defines the following set of HTTP interfaces:

/translate

Translates document immediately.
Limitations: size, frequency, number of queued documents, format, …

Input (POST):

- **auth** – authentication data (future, ignored)
  - Username/password, authentication key, …
  - Currently ignored
- **src** – source language (future, optional)
  - Follows ISO 639-1 language code
  - Currently only English ('en') is supported
  - Default: 'en'
- **dst** – destination language (required)
  - Follows ISO 639-1 language code
  - Currently supported languages:
    - Bulgarian ('bg')
    - Czech ('cs')
    - German ('de')
    - Greek ('el')
    - Croatian ('hr')
    - Italian ('it')
    - Dutch ('nl')
    - Polish ('pl')
    - Portuguese ('pt')
    - Russian ('ru')
    - Chinese ('zh')
  - No default
- **domain** – domain (optional)
  - Currently only 'informal' domain is supported
  - Default: 'informal'
- **type** – document format (optional)
  - Text, subtitle (WebVTT, srt, DFXP, …), HTML, docx, pptx, PDF, …
  - Follows MIME type specification
  - Currently the following formats are supported:
    - text ('text/plain')
    - subtitle ('text/vtt')
      Subtitle format is converted to and returned as WebVTT.
    - HTML ('text/html')
      Text inside <pre> and comment blocks is not translated.
    - Microsoft Open XML documents:
      - MS Word docx (' application/vnd.openxmlformats-officedocument.wordprocessingml.document ')
      - MS PowerPoint pptx (' application/vnd.openxmlformats-officedocument.presentationml.presentation')
      - MS Excel xlsx (' application/vnd.openxmlformats-officedocument.spreadsheetml.sheet')
  - Default: 'text/plain'
- **data** – document (data, required)
  - Should be encoded as UTF-8

Output (raw data):

- Translated document
    - Encoded as UTF-8

Errors:

- 500 Internal Server Error
    - All other possible errors
- 503 Service Unavailable
    - Translation service unavailable
- 400 Bad Request
    - Target language not present
    - Invalid target language specified
    - Zero length or no input data

Example – translate text in file sample.txt from English to German:
$ curl -F dst=de -F type=text/plain -F data=@sample.txt -u user:password http://matterhorn.ijs.si/trans/translate
200 OK
...translated text...

Example – translate subtitle sample.vtt from English to Italian:
$ curl -F dst=it -F type=text/vtt -F data=@sample.vtt -u user:password http://matterhorn.ijs.si/trans/translate
200 OK
...translated subtitle...

/ingest

Ingests a document into a queue for postponed translation.
Limitations: size, number of queued documents, format, …

Input (POST):

- **auth** – authentication data (future, ignored)
    - Username/password, authentication key, …
    - Currently ignored
- **src** – source language (future, optional)
    - Follows ISO 639-1 language code
    - Currently only English ('en') is supported
    - Default: 'en'
- **dst** – destination language (required)
    - Follows ISO 639-1 language code
    - Currently supported languages:
        - Bulgarian ('bg')
        - Czech ('cs')
        - German ('de')
        - Greek ('el')
        - Croatian ('hr')
        - Italian ('it')
        - Dutch ('nl')
        - Polish ('pl')
        - Portuguese ('pt')
        - Russian ('ru')
        - Chinese ('zh')
    - No default
- **domain** – domain (optional)
    - Currently only 'informal' domain is supported
    - Default: 'informal'
- **type** – document format (optional)
    - Text, subtitle (WebVTT, srt, DFXP, …), HTML, docx, pptx, PDF, …

    - Follows MIME type specification
    - Currently the following formats are supported:
        - text ('text/plain')
        - subtitle ('text/vtt')
          Subtitle format is converted to and returned as WebVTT.
        - HTML ('text/html')
          Text inside <pre> and comment blocks is not translated.
        - Microsoft Open XML documents:
            - MS Word docx (' application/vnd.openxmlformats-officedocument.wordprocessingml.document ')
            - MS PowerPoint pptx (' application/vnd.openxmlformats-officedocument.presentationml.presentation')
            - MS Excel xlsx (' application/vnd.openxmlformats-officedocument.spreadsheetml.sheet')
    - Default: 'text/plain'
- **data** – document (data, required)
    - Should be encoded as UTF-8
- **prio** – document priority (for queue, optional)
    - Integer value between 1 and 9
    - Default: 5
- **callback** – callback URL (optional)
    - HTTP or e-mail notification, executed when translation is done
    - Follows URI specification
    - Currently only e-mail notification ('mailto:user@domain.com') is supported
    - No default

Output (JSON):

- Document ID
- Status (queued)
- Source language (en)
- Destination language
- Domain (informal)
- Document type
- Queue entered date
- Translation started date (null)
- Finished date (null)
- Translation status (null)
- Callback URL
- Priority
- Document size
- Queue position

Errors:

- 500 Internal Server Error
    - Insert into database failed
    - All other possible errors
- 503 Service Unavailable
    - Translation service unavailable
- 400 Bad Request
    - Target language not present
    - Invalid target language specified
    - Zero length or no input data

Example – put text in file sample.txt into translation queue from English to German:
$ curl -F dst=de -F callback=mailto:user@domain.com -F data=@sample.txt -u user:password http://matterhorn.ijs.si/trans/ingest
{

```
  "id": 2,
  "status": "queued",
  "src_lang": "en",
  "dst_lang": "de",
  "domain": "informal",
  "type": "text/plain",
  "enter_date": "2017-10-31T12:31:53.052",
  "trans_date": null,
  "finish_date": null,
  "trans_status": null,
  "callback": "mailto:user@domain.com",
  "prio": 5,
  "size": 1536,
  "que_pos": 1
}
```


When queued text has been processed, the following e-mail arrives:
From: Translation Service <user@ijs.si>
Subject: Translated request id: 2 (en-de, plain text)

Translated request id: 2
Source language: en
Target language: de
Document type: text/plain
Document size: 1536
Entered in queue: 2017-10-31 12:31:53.052047
Translation finished: 2017-10-31 12:50:45.183930
Translation status: success

Attached to the e-mail are two files:
- Original input text (random file name with an extension '.en.txt')
- Translated output text (random file name with an extension '.en-de.txt')

/ingest_status

Returns status of queued document.

Input (GET):
- **auth** – authentication data (future, ignored)
  - Username/password, authentication key, …
  - Currently ignored
- **id** – document ID (required)
  - as returned from /ingest API

Output (JSON):
- Document ID
- Status (queued, finished or failed)
- Source language (en)
- Destination language
- Domain (informal)
- Document type
- Queue entered date
- Translation started date
- Finished date
- Translation status
- Callback URL

- Priority
- Document size
- Queue position

Errors:
- 500 Internal Server Error
  - All other possible errors
- 400 Bad Request
  - No document id specified or id is empty
- 404 Not Found
  - Document id <n> not found

Example – check status of document in translation queue:
$ curl -u user:password http://matterhorn.ijs.si/trans/ingest_status?id=2

```
{
  "id": 2,
  "status": "finished",
  "src_lang": "en",
  "dst_lang": "de",
  "domain": "informal",
  "type": "text/plain",
  "enter_date": "2017-10-31T12:31:53.052",
  "trans_date": "2017-10-31T12:50:31.181",
  "finish_date": "2017-10-31T12:50:45.183",
  "trans_status": "success",
  "callback": "mailto:user@domain.com",
  "prio": 5,
  "size": 1536,
  "que_pos": null
}
```


/ingest_control

Executes action (get, modify or delete) on queued document. You can get current status of queued document with /ingest_status API.

Input (GET):
- **auth** – authentication data (future, ignored)
  - Username/password, authentication key, …
  - Currently ignored
- **id** – document ID (required)
  - as returned from /ingest API
- **action** – action to be executed on document ID (required)
  - Currently supported actions:
    - **get**
      Transfers translated document as raw data.
      Usable for example when no callback was specified.
    - **modify**
      Modifies specified document ID in queue.
      You can modify src, dst, domain, type, prio and callback parameters. This is reasonable only before the document has been processed.
    - **delete**
      Deletes specified document ID from queue.
  - No default
- **src** – source language (future, optional for modify action)
  - Follows ISO 639-1 language code
  - Currently only English ('en') is supported

    - No default
- **dst** – destination language (optional for modify action)
    - Follows ISO 639-1 language code
    - Currently supported languages:
        - Bulgarian ('bg')
        - Czech ('cs')
        - German ('de')
        - Greek ('el')
        - Croatian ('hr')
        - Italian ('it')
        - Dutch ('nl')
        - Polish ('pl')
        - Portuguese ('pt')
        - Russian ('ru')
        - Chinese ('zh')
    - No default
- **domain** – domain (optional for modify action)
    - Currently only 'informal' domain is supported
    - No default
- **type** – document format (optional for modify action)
    - Text, subtitle (WebVTT, srt, DFXP, …), HTML, docx, pptx, PDF, …
    - Follows MIME type specification
    - Currently the following formats are supported:
        - text ('text/plain')
        - subtitle ('text/vtt')
          Subtitle format is converted to and returned as WebVTT.
        - HTML ('text/html')
          Text inside <pre> and comment blocks is not translated.
        - Microsoft Open XML documents:
            - MS Word docx (' application/vnd.openxmlformats-officedocument.wordprocessingml.document ')
            - MS PowerPoint pptx (' application/vnd.openxmlformats-officedocument.presentationml.presentation')
            - MS Excel xlsx (' application/vnd.openxmlformats-officedocument.spreadsheetml.sheet')
    - No default
- **prio** – document priority (optional for modify action)
    - Integer value between 1 and 9
    - No default
- **callback** – callback URL (optional for modify action)
    - HTTP or e-mail notification, executed when translation is done
    - Follows URI specification
    - Currently only e-mail notification ('mailto:user@domain.com') is supported
    - No default

Output (raw data):

- Translated document
    - Encoded as UTF-8

Output (JSON):

- Document ID
- Status (modified, deleted or failed)

Errors:

- 500 Internal Server Error
    - Update database failed
    - All other possible errors
- 400 Bad Request

    - No document id specified or id is empty
    - No action specified or action is empty
    - No valid action specified
    - Document id <n> has not been translated yet
- 404 Not Found
    - Document id <n> not found

Example – get the translated document from the translation queue:
$ curl -u user:password http://matterhorn.ijs.si/trans/ingest_control?id=2&action=get
200 OK
...translated text...

Example – modify the priority of the document in the translation queue:
$ curl -u user:password http://matterhorn.ijs.si/trans/ingest_control?id=2&action=modify&prio=9

{
  "id": 2,
  "status": "modified"
}

Example – delete the document in the translation queue:
$ curl -u user:password http://matterhorn.ijs.si/trans/ingest_control?id=2&action=delete

{
  "id": 2,
  "status": "deleted"
}

/system_status

Returns system status.

Output (JSON):
- API version
- System status (available or unavailable)
- List of available language pairs
- Concurrency (number of possible concurrent NMT cores/language pairs)
- Current queue size

Errors:
- 500 Internal Server Error
    - All other possible errors

Example – get the system status:
$ curl -u user:password http://matterhorn.ijs.si/trans/system_status

{
  "API_version": "1.0",
  "status": "available",
  "lang_pairs": ["en-bg", "en-cs", "en-de", "en-el", "en-hr", "en-it", "en-nl", "en-pl", "en-pt", "en-ru", "en-zh"],
  "concurrency": 2,
  "queue_size": 0
}

/identify_language (future)

Identify document language (not implemented yet).

Input (POST):

- **auth** – authentication data (future, ignored)
  - Username/password, authentication key, …
  - Currently ignored
- **type** – document format (optional)
  - Text, subtitle (WebVTT, srt, DFXP, …), HTML, docx, pptx, PDF, …
  - Follows MIME type specification
  - Currently the following formats are supported:
    - text ('text/plain')
    - subtitle ('text/vtt')
      Subtitle format is converted to and returned as WebVTT.
    - HTML ('text/html')
      Text inside <pre> and comment blocks is not translated.
    - Microsoft Open XML documents:
      - MS Word docx (' application/vnd.openxmlformats-officedocument.wordprocessingml.document ')
      - MS PowerPoint pptx (' application/vnd.openxmlformats-officedocument.presentationml.presentation')
      - MS Excel xlsx (' application/vnd.openxmlformats-officedocument.spreadsheetml.sheet')
  - Default: 'text/plain'
- **data** – document (data, required)
  - Should be encoded as UTF-8

Output (JSON):
- Detected document language (currently hardcoded to 'en')

Errors:
- 500 Internal Server Error
  - All other possible errors
- 400 Bad Request
  - Zero length or no input data

Example – identify document language:
$ curl -F type=text/plain -F data=@sample.txt -u user:password
http://matterhorn.ijs.si/trans/identify_language

{
  "lang": "en"
}