MetaDoc Client Documentation

Version 0.1.0

Bjørnar Grip Fjær

July 2, 2010

Contents

1 Introduction

MetaDoc is created as a way to securely transport data between a server and sites. It is created to improve the flow of information between High Performance Computing (HPC) sites and Uninett Sigma [1].

MetaDoc consists of a client, running at the site, and a server running at the Metacenter. MetaDoc takes care of authenticating the client on the server, packing and unpacking the data to and from Extensible Markup Language (XML), and transporting the data securely between client and server.

2 Using the MetaDoc client

Usage of the MetaDoc client is done mainly through the use of `main.py`. `main.py` takes care of sending and retrieving data to and from the server, caching any data that could not be sent, and validating XML data recieved.

When `main.py` should send data to the server, a custom function that should populate the data to be sent is called, so that each site can customize the way data is gathered on the site.

When `main.py` recieves data from the server, it calls a custom function based on the data recieved, where each site can define what should be done with the recieved data.

Section 2.5 explains what functions are called and how they should handle the data.

2.1 Configuration

The MetaDoc client uses a configuration file `metadoc.conf`. This file *must* be placed in the same folder as the client itself.

The configuration is in INI format and defines a section named `MetaDoc` which must contain the following parameters:

host **baseurl** for the MetaDoc Server that the client should communicate with.

key Absolute path to the clients SSL certificate key file. This should be the private key for **cert**, and is used to encrypt data passed to the server.

cert Absolute path to the clients SSL certificate file. This is used to authenticate the client with the server. See section 3.2 for more information.

site_name The name of the site the client is running on.

The following parameters *may* be defined:

trailing_slash Whether the client should append a trailing slash at the end of URLs used to connect to the server. At the moment, this should be set to `True`.

valid Initially set to `False` when `main.py` creates a sample configuration file. This is set to avoid running the client without being properly configured. If this value is present, it *must* be set to `True` or `yes`.

`main.py` will create a initial configuration file if one is missing when it starts. This can be used as a basis for a configuration file.

2.2 Handles

`main.py` takes handles that tells the script what information to send or retrieve to or from the server. All handles can be mixed together *except* for handles that override each other. Handles that overrides others are explicitly stated below.

`main.py` takes the following handles:

-h, –help Displays a short help message explaning the handles that may be passed to `main.py`. Overrides any other handles.

-v, –verbose Verbose mode. Prints information about progress and information sent and recieved between client and server.

-q, –quiet Quiet mode. Prints nothing unless the program fails. Overrides **-v**, **–verbose**.

-l <log level>, –log-level=<log level> Sets the log level for the program. See section 2.3 for more information about what is logged at different levels.

-n, –no-cache Prevents the client from sending any cached data. For more information about caching, see section 2.4.

-e Sends event data from client to server.

-c Sends configuration data from client to server.

-s Sends software data from client to server.

-u Retrieves user data from the server.

-a Retrieves allocation data from server.

-p Retrieves project data from server.

–dry-run Does a dry run, not sending any data to server. Does not retrieve cached data, and does not save any data to cache. Should be run with verbose to see data that would be sent.

2.3 Logging

The client logs data to `/var/log/mapi/`. The folder must be read and writable for the user that runs the client in order for the client to run. The client creates a new log file each day it is run, with the name `metadoc.client.YYYY-mm-dd.log`.

The client has five different logging levels. The list below gives an overview of what is logged at the different levels. The higher items in the list contain everything below as well, so that with a log level set to **error** will also contain **critical** logging.

debug Debugging information, used for development and error checking.

info Information about what is happening during execution, such as items sent or recieved to/from the server.

warning Warnings occouring during execution, mainly problems that will not cause a failure but that should be addressed.

error Errors that cause partial failure of the execution, such as being unable to connect to the server.

critical Critical failures that causes the execution to halt, or errors in the program code itself.

The log level defaults to the lowest possible, so everything will be logged if nothing is set.

2.4 Caching

The client caches data to `/var/cache/mapi/`. Files are named after the data type that is cached in each file. The user running the client must have read and write access to the folder in order for the client to run.

The client will cache any information that is not accepted by the server, *unless* the server returns a reciept for the information that marks the information as invalid or malformed in some way, such that the information will not be accepted if resent at a later date. See section 8 for more information.

Data the client sends may be marked so that it will not resend any cached data when the client is run with the same handle. This is mainly for use for full updates, such as software and configuration, where any cached data would be outdated or duplicates if sent together with a new run.

If the **-n** or –**no-cache** handles are passed, the script will ignore any cached data completely and run as if it didn't exist. The cached data will then be processed on the next run where **-n** or –**no-cache** is not passed.

2.5 Customizing the MetaDoc client

The MetaDoc client calls a specified function based on the data passed between client and server. Only one type of data should be sent per document, and both the client and the server only checks for the expected type of data in the recieved XML document.

Each data type has a named container element within the XML document, which there should only be one of per document. If a list of data is passed between server and client, the list should be contained within a container element. The name of the container element is used in naming modules and classes in order to ease readability of code.

The naming of the class that handles the data passed between server and client on the client side depends on whether data is passed from client to server, or the other way around. For information about the classes and functions used when sending data to the server, see section 2.5.1, and for data recieved from the server, see section 2.5.2.

2.5.1 Sending data

Data sent from the client to the server must first be populated. Because there is not always a standard way to populate this data on every site, a custom class is created. This class should be found under `custom.site<name>.Site<Name>`, where `<name>` is the name of the container element in the XML, and `<Name>` is the name capitalized (e.g. `config` would use `custom.siteconfig.SiteConfig`). This class should inherit from `custom.abstract.MetaOutput`.

When the client is ready to fetch these items, the function `populate()` on this class will be called. This function should gather any information to be sent, create elements found in `<name>.definition.<Name>.legal_element_types` for this information and place these items in `self.items`.

The client will then take care of packing data to XML, sending the data to the server, processing the receipt returned from the server and caching any data that was not accepted by the server. For more information on caching, see section 2.4.

2.5.2 Recieving data

Data that is recieved from the server must be processed. The client will take care of fetching the data from the server, unpacking the XML and creating objects based on the type of data recieved. Once this is done, a function called `process()` will be called on the class `custom.update<name>.Update<Name>`. When this function is called, the object's `self.items` should be a list of `<name>.definition.<Name>` objects.

Examples for producing files similar to the ones now in use based on information transferred through MetaDoc is given in `doc/examples/custom/`.

3 MetaDoc Server API

The MetaDoc server implements a REST-like API, however, there are certain differences from REST noted in section 3.3.

When the client performs a GET request on an availible URL, the server should return an XML document, or a HTTP status code refering to an error. The XML document should follow the MetaDoc DTD [2]. Each URL only returns data from the requested data type. This means that a request to **baseurl/allocations/** will return an MetaDoc XML document containing only an `<allocations>` directly on the `<MetaDoc>` root, with `<all_entry>` tags as children of `<allocations>`. The client should disregard any information outside `<allocations>` when connecting to **baseurl/allocations/**.

In order to send data to the server, the client performs a POST request, with the POST data variable `metadoc` containing a MetaDoc XML document. The server will only accept data from the data type specified in the URL, and will disregard any other information. This means that a POST to **baseurl/events/** should be a MetaDoc XML document containing a `<events>` tag directly on the `<MetaDoc>` root, with any number of `<resourceUp>` and `<resourceDown>` tags as children of `<events>`.

When this data is sent to the server, the server should return a MetaDoc XML document containing a `<receipts>` tag, with a `<r_entry>` tag for each element recieved that has an **id**-attribute.

3.1 Availible URLs

baseurl/allocations/ Retrieves a list of allocations/quotas relevant to the client

baseurl/users/ Retrieves a list of users for the client

baseurl/projects/ Retrieves a list of projects relevant to the client

baseurl/config/ Sends system configuration to server

baseurl/events/ Sends site events to the server

baseurl/software/ Sends system software to server

3.2 Authentication

The site uses SSL certificates to authenticate the client. In order for the site to be authenticated properly, the server must be aware of the client's certificate already, and the correct owner of the certificate must be saved on the server. This *must* be the same as the value for `site_name` set in the MetaDoc configuration (see section 2.1).

3.3 Differences from REST

There are certain differences in the API compared to the REST specification. The MetaDoc Server API makes use of HTTP POST where HTTP PUT should be used in accordance with REST. This is due to limitations in standard Python libraries.

Because the access the MetaDoc Server API gives to the client is limited, this change does not prohibit any other functionality.

No address currently supports both adding and retrieving data. This is not a limitation in the system itself, but a matter of authorative sources of information.

3.4 Server HTTP responses

The server makes use of HTTP status codes to identify what error has occoured if the server is unable or unwilling to process the request from the server.

If the client does not send a SSL certificate, sends a sertificate unknown to the server, attempts to get information about sites not identified with the certificate, or attempts to send information that identifies as another site, the server returns a "403 Forbidden" status code.

If the server fails to process the request, it will return a "500 Server Error" status code. This does *not* include errors on the document that results in returning receipts.

4 XML document

The XML document should follow the form described in the MetaDoc DTD [2]. Below certain conventions used in the XML build is explained. Any alterations to the DTD should follow these conventions in order for the client and server to continue functioning normally.

4.1 Document build

Any type of information sent should only create one direct child of the root element, `<MetaDoc>`. This means that when lists of information is sent, the list elements should be placed within a container element, and *not* directly in the root element. The container element should have an self explanatory name about the information passed.

An example is that `<user_entry>` elements are placed within a `<users>` element. Here `<users>` is considered the container element, and there should only be one of them in each MetaDoc XML document. `<users>` may contain any number of `<user_entry>` elements.

4.2 Dates

All dates in the document should be on the form specified by RFC3339 [3]. The `utils` module provides a function `date_to_rfc3339` that takes a `datetime.datetime` object and returns a string on RFC3339 form. It also provides a function `rfc3330_to_date` which will return a `datetime.datetime` object from a proper RFC3339 string, or **False** if the string is not a correct RFC3339 date.

4.3 Special attributes

The **id** attribute of elements have a special function in MetaDoc. This attribute is used to identify the object when recieving receipts from the server whether elements have been added. The attribute is *not* saved in caching to avoid duplicate **id**s when resending cached data together with new data. If you want to give elements a special identifier that should be saved, it must be called something other than **id**.

5 Useful classes and modules

The MetaDoc client defines a series of classes and modules that are used to define the XML data passed between client and server.

`MetaElement` refers to `metaelement.MetaElement`, and `MetaDoc` refers to `metaelement.MetaElement`.

5.1 MetaDoc

The class `MetaDoc` is used to define the document itself. It provides functionality to alter the document structure, find elements within the document and generate XML data from the document. It contains a series of `MetaElement` sub classes that defines the content within the document.

5.2 MetaElement

`MetaElement` is used to define content within the document. The class is ment to be a parent class for other classes that defines particular elements. An element is equal to an XML tag, such as `<users>` or `<resourceUp>` in a MetaDoc XML document. An instance of such a sub class is used to define a particular tag in the XML document, with attributes and content.

5.2.1 Class variables

Each `MetaElement` sub class should define a class variable **xml_tag_name**, that should be a string containing the name of the XML tag the class describes. For a sub class defining the `<users>` tag, this should be set to "users".

A `MetaElement` sub class that defines a main container element, that is, an element that is placed as a direct child of the root node `<MetaDoc>` in the XML document, should also define a class variable **url** that is a string containing the particular part of the URL that is used to send or retrieve information for the data type passed. If the type of data is a list of users, and it should retrieve the list of users from the url **/baseurl/users/**, the **url** class variable should be set to "users".

The classes that define a main container element should also define either the class variable **update_handler** or **site_handler**, depending on whether the data is ment to be recieved from the server or sent from the client, respectively.

update_handler should be a sub class of `custom.abstract.MetaInput`. This class should be placed in `custom.update<name>.Update<Name>`, so for users this would be `custom.updateusers.UpdateUsers`. When data of the type defined by the `MetaElement` sub class is recieved, an instance of **update_handler** will be created, and the instance's **self.items** will be populated with a list of `MetaElement` sub classes. Then the **update_handler**'s `process()` function will be called. Normally, **self.items** should be of length 1.

site_handler should be a sub class of `custom.abstract.MetaOutput`. This class should be placed in `custom.site<name>.Site<Name>`, so for events this would be `custom.siteevents.SiteEvents`. When the script is called to send data of the type defined by the `MetaElement` sub class, an instance of **site_handler** is created, and the instance's `populate()` function is called. `populate()` should populate the instance's **self.items** with a list of `MetaElement` sub classes. When `populate()` is done, **self.items** is added to the `MetaElement` sub class instance's **self.sub_elements**.

5.2.2 Allowed sub elements

There are certain restrictions on what classes can be placed within a `MetaElement` sub class instance's **self.sub_elements**, because not every XML tag can have any other XML tag as children. A `MetaElement` sub class therefor defines a **self.legal_element_types**. This should be a list of `MetaElement` sub classes that are allowed to be children of the XML current `MetaElement` sub class.

As an example, if `Users` is a `MetaElement` sub class defining the `<users>` XML tag, and `UserEntry` is a `MetaElement` sub class defining the `<user_entry>` XML tag, which can be a child of `<users>` in the XML document, a `Users` instance would have `UserEntry` in it's **self.legal_element_types**.

5.2.3 Tag attributes

A tag may have attributes set. These should be defined in the `MetaElement` sub class' `__init__()` function. They *must* have the same name as the attribute has in the XML document. If an attribute is optional in the XML element, it should also be optional in `__init__()`.

The sub class should in it's `__init__()` function figure out what attributes are availible and which are not, and pass these on to the `MetaElement` `__init__()` function through `super()`.

5.3 UniqueID

The `utils` module provides a class called `UniqueID` that can provide a unique identifier to objects passed through the function `get_id()`. This should be used for entries passed from client to server to set as the **id** attribute so that the server can properly identify the entry when returning a receipt.

5.4 Examples

5.4.1 Connection figure

Figure 1 shows an example of how these connections work. Here the definition of projects is shown, with connections to the XML document, DTD, server URL and **update_handler**.

5.4.2 Script example

Below a small example of transferring posts from client to server is shown.

Script 1: posts.definition

```
1  import metaelement
2  from custom.siteposts import SitePosts
3  from posts.entries import PostEntry
4
5  class Posts(metaelement.MetaElement):
6      xml_tag_name = "posts"
7      site_handler = SitePosts
8      url = "posts"
9
10     def __init__(self):
11         super(Posts, self).__init__(Posts.xml_tag_name)
12         self.legal_element_types = (PostEntry,)
```

Script 2: posts.entries

```
1  import metaelement
2  import re
3  from custom.siteposts import SitePosts
```

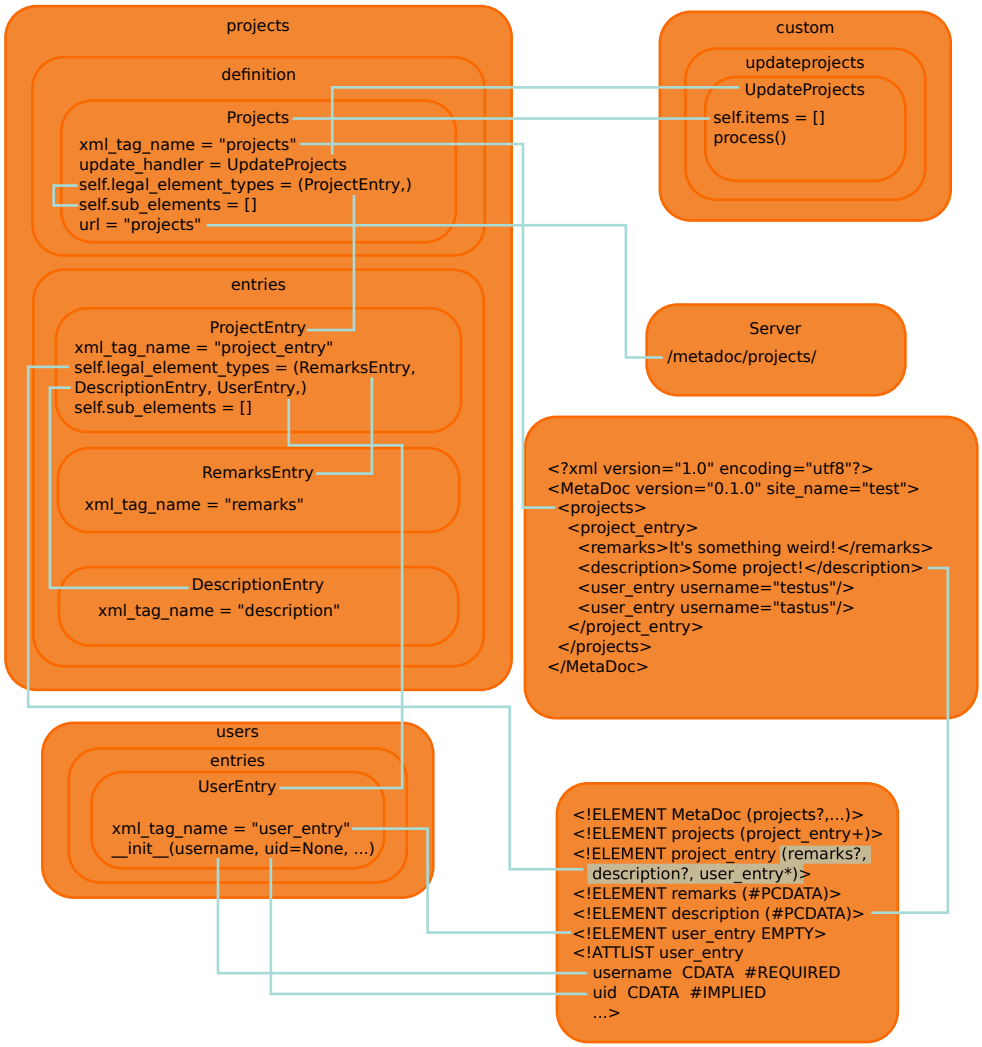

Figure 1: Example of how projects data is defined and connection between classes used in definition and processing of project data.

```
4   from utils import UniqueID()
5
6   class PostEntry(metaelement.MetaElement):
7       xml_tag_name = "post_entry"
8
9       def __init__(self, title, author):
10          unique_id = UniqueID()
11          attributes = {
12              'title': title,
13              'author': author,
14              'id': unique_id.get_id(),
15          }
16          super(PostEntry, self).__init__(PostEntry.xml_tag_name, attributes)
17      def set_content(self, content):
18          """ Sets the content of the post. """
19          self.text = content
20      def clean_author(self, author):
21          """ Makes sure the author's name is in lower case. """
22          return author.lower()
```

Script 3: custom.siteposts

```
from custom.abstract import MetaOutput
from posts.entries import PostEntry

class SitePosts(MetaOutput):
    def populate(self):
        """ Adds a couple of posts that should be sent. """
        post_one = PostEntry(title="We're making progress!", author="bjornarg")
        post_one.set_content("""This text will be in post_entry MetaElement
instance's self.text!""")
        self.items.append(post_one)
        post_two = PostEntry(title="Cleaning attributes" author="ProperAuthor")
        post_two = "We've made sure that author is sent in lower case."
        self.items.append(post_two)
```

Script 4: XML Example

```
<?xml version="1.0" encoding="utf-8"?>
<MetaDoc version="0.0.0" site_name="example">
  <posts>
    <post_entry title="We're making progress!" author="bjornarg" id="_1">
      This text will be in post_entry MetaElement instance's self.text!
    </post_entry>
    <post_entry title="Cleaning attributes" author="properauthor" id="_2">
      We've made sure that author is sent in lower case.
    </post_entry>
  </posts>
</MetaDoc>
```

6 Extending MetaDoc

To extend MetaDoc to send more data, the following is needed. Each step is explained in more detail below. Certain restrictions is set on how the XML document should be formed. See section 4 for more information.

- Definition of data to be sent in the MetaDoc DTD.
- A definition file explaining the data on the client.
- An entries file, explaining any entries allowed in the data on the client.
- A `MetaInput` or `MetaOutput` sub class that should handle data, depending on whether data is recieved or sent to or from the server, respectively.
- Adding a handle to `main.py` that will activate the data type.
- Configuring the server to send or recieve the intended data.

Figure 1 shows an example of how the data for projects is defined, and the connection between classes that are used when data about projects is recieved from the server.

6.1 Altering DTD

The XML document follows certain conventions that should be followed when extending the DTD. These conventions are explained in more detail in section 4.

Before you alter the DTD you should know exactly what data should be sent. Create an `<!ELEMENT>` with a descriptive name of the data sent. As an example, `<users>` is used for a list of users.

Add any attributes necessary to describe the set of data. If the data is a list of entries, such as a list of users, create an `<!ELEMENT>` as a possible sub-element that contains the information about each entry. Any short information about the entry should be placed in attributes of the entry. If there is more information, such as information that could be several sentences or lines, it should not be placed as an attribute. This information should be placed inside the element itself. If there are several types of long information for each entry, create a descriptive `<!ELEMENT>` for each as a sub-element of each entry to contain the text. Otherwise the text may be placed directly inside the entry element itself.

6.2 Defining the data on the client

Add a module to the client with the name of the main element. Create a file `definition.py` inside this module. `definition.py` should define the main element with a sub class of `metaelement.MetaElement`.

Create a file `entries.py` inside the same module. This file should contain definitions of each entry, and potential sub elements for each entry, as a sub class of `metaelement.MetaElement`.

Add the entry class(es) to `self.legal_element_types` of the `metaelement.MetaElement` sub class defining the main element.

`metaelement.MetaElement` sub classes may define a `clean_<attribute name>()` for each attribute on the element. This method will recieve the attribute value, and should return the attribute value after any potential cleaning is done on it. Please note that *all* attribute values *must* be strings, so if any value set as an attribute might be set as anything other than a string, the clean-function is the place to convert it.

See section 5 for more information on how to build these classes.

6.3 Custom client handles

If the data is to be sent from client to server, create a module `custom/site<main element name>.py` that contains a sub class of `custom.abstract.MetaOutput`. This class should define a method `populate()` which gathers the information to be sent from the site and appends an instance of a entry-class, as defined in section 6.2, to **self.items** for each entry.

Section 2.5.1 and section 5 contain more information on how to define these classes.

If the data is sent from server to client, a module called `custom/update<main element name>.py` should be created that contains a sub class of `custom.abstract.MetaInput`. This class should define a method `process()` that processes any recieved data in **self.items**.

6.4 Versioning

MetaDoc passes a **version** attribute on it's root element, `<MetaDoc>` when sending information between client and server. This version is a number on the form "`X.Y.Z`", where `X`, `Y` and `Z` are numbers. Changes made to each number indicate different levels of breakage.

When `X` is changed, changes are made such that the current information passed is changed in some way. This may be changes to the DTD where any of the currently passed information is affected (addition/removal of attributes, changes to how attribute values are presented or should be parsed, addition/removal of sub-elements). If the client or server encounters a document with a different value of `X` in the version number, it should *not* accept the data, as it cannot be sure it will handle it correctly.

Changes to `Y` indicates a change that does *not* change the current behaviour in any way, but instances where new information might be passed. When the client or server encounters a document with a different value of `Y` it should log a warning, but otherwise proceed as normally.

`Z` is currently not used for anything, but is present for potential usability in the future. Differences in `Z` should be logged as debug information.

7 Information flow

The client will always be the initiator in either requesting data from the server or sending data.

Figure 4 shows how information passes when a client requests a user list. The steps are as follows:

1. `main.py` is run with the handle **`-u`**, which will check `users.definition.Users` for which URL to access on the server to retrieve the information.
2. A request is sent to the server to retrieve the information.
3. The server creates an XML document with a list of users for the site.

   The XML document sent is defined by the MetaDoc DTD.
4. `main.py` recieves the XML document, checks that it is valid according to the DTD.
5. If the XML document passes validation, an instance of `users.definition.Users` is created, and an instance of `users.entries.UserEntry` is created for each `<user_entry>` in the XML document.

   `users.definition.Users` and `users.entries.UserEntry` may validate attributes and refuse to create any elements where attribute validation does not pass.
6. The list of validated `users.entries.UserEntry` instances is placed within `self.items` for an `custom.updateusers.UpdateUsers` instance, and the `process()` function is called for the processing of the user list.

7.1 Validation

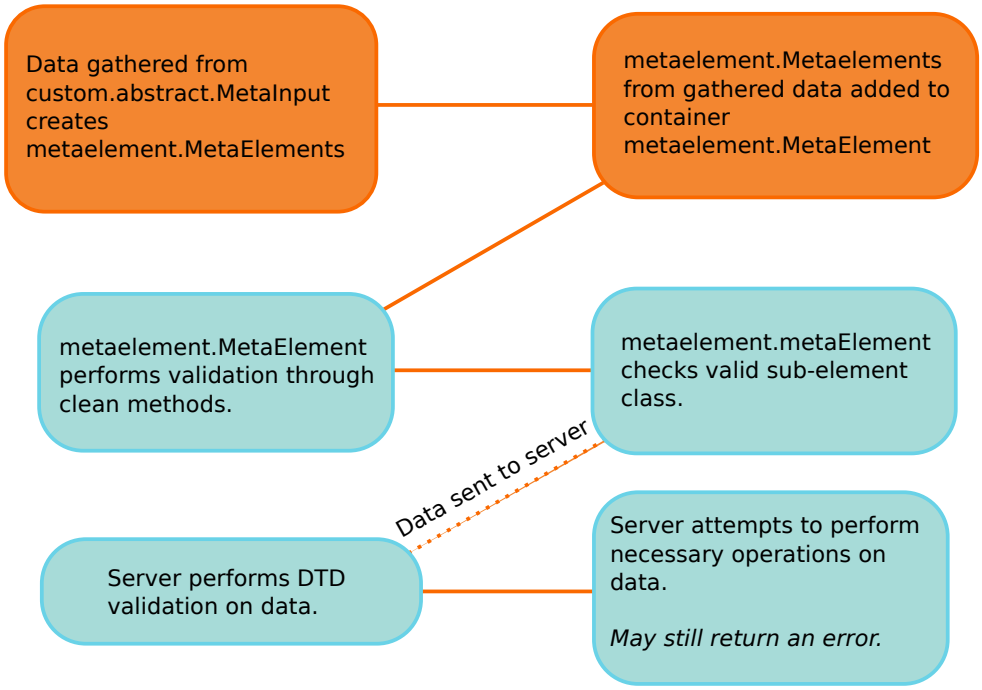

Figure 2: Shows validation procedures when data is passed from client to server

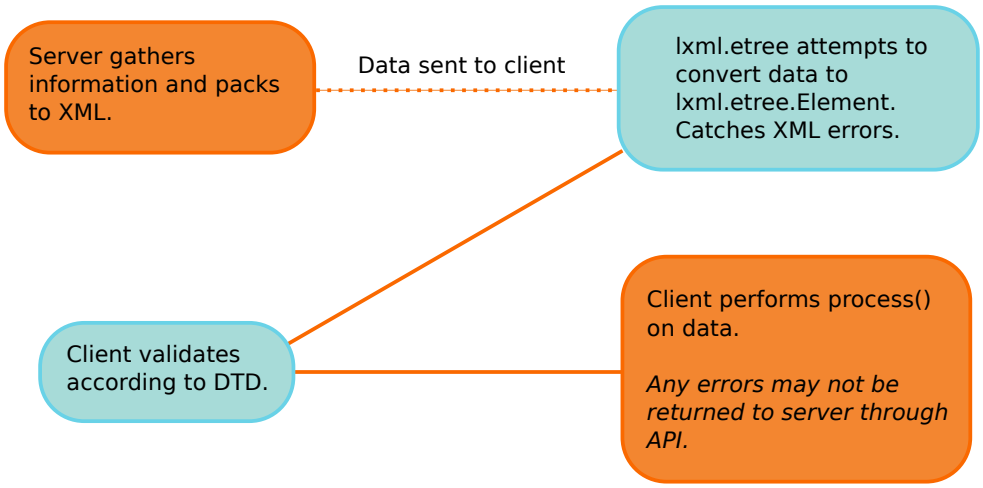

Figure 3: Shows validation procedures when data is passed from server to client

8 Errors

The server returns a `<receipt>` containing an `<r_entry>` for each element parsed. The `<r_entry>` has the required attributes **id** and **code**, containing the ID of the element and the error code, respectively. It may also contain an attribute **note** with a short note explaining the error if extra information is availible. The `<r_entry>` tag might also contain text with a longer message, if more information is needed about the error.

8.1 Document errors

In the special case where there are problems with the document itself, such as XML errors or the document not passing DTD verification, the `<r_entry>` **id** attribute will be set to `0` (zero), reffering to the document itself.

References

[1] Austad, Henrik, *Improving the Information Flow within the Metacenter*, `http://www.austad.us/metadoc/improvingFlow.pdf`

[2] *MetaDoc Document Type Definition*, `http://bjornar.me/metadoc/MetaDoc.dtd`

[3] *RFC3339*, `http://www.ietf.org/rfc/rfc3339.txt`

A List of errors

Table 1: Error codes recieved from server

Error code	Meaning	Extra notes
1000	No errors	
2000	Error with the XML data	
2001	Missing attribute	Missing attribute should be returned as a note.
5000	Database error	
5001	MySQL database error	Note should contain the MySQL error code, and the message the MySQL error message
6001	Another element error	Another element in the same set has been rejected, and this type of set will not accept any elements if any element contains an error

B Information flow

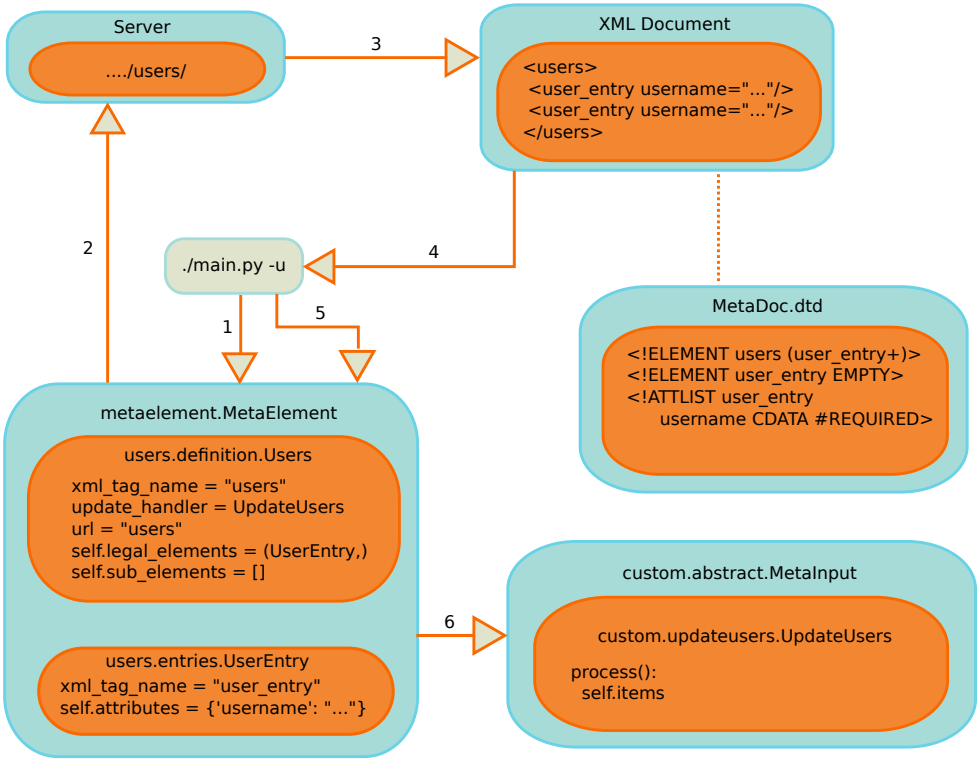

Figure 4: Information flow when requesting user list with MetaDoc

C Included examples

`doc/examples/` includes a set of examples for using MetaDoc. Below is a list of the included examples and what they do.

cli/event.py A command line interface for adding events. Should be placed inside `client/` when run. See `event.py --help` for usage.

custom/updateusers.py Custom function for converting recieved user data into a passwd/shadow file.

custom/updateprojects.py Custom function for converting recieved project data to a project user file.

custom/updateallocations.py Custom function for converting recieved allocation data into a quota file for projects.