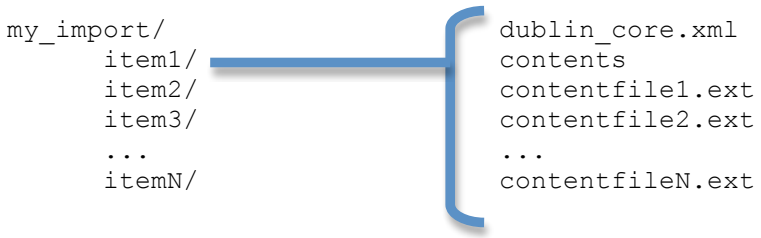
DSpace Batch Import Format

The batch import system for DSpace is a simple but primitive method for importing multiple items into a DSpace repository. Typically a custom script will need to be created which migrates data from a source into this DSpace batch import format. Possible source formats are Excel spreadsheets, Access databases, other websites, or a flat file system. The custom script will produce a set of directories and files that correspond with the format detailed in this document.

The Format:

First create a working directory, like "`my_import`". For each item you create a new directory inside "`my_import`". Inside the directories you will need to place the descriptive metadata and all the content files to be included with each item item. Here is an example of the directory structure for a single item:



The `dublin_core.xml` contains the descriptive metadata about the item. This well-formed XML file is simple a list of `<dcvalue>` elements, each with its Dublin Core element, qualifier and value. You will need to check with the repository administrator about what Dublin Core Element and Qualifiers are available, and which values should be placed in these fields. Here is an example metadata file:

```
<?xml version="1.0" encoding="UTF-8"?>
<dublin_core>
    <dcvalue element="contributor" qualifier="author">Public, John Q.</dcvalue>
    <dcvalue element="language" qualifier="iso">en</dcvalue>
    <dcvalue element="subject" qualifier="none">Technology</dcvalue>
    <dcvalue element="title" qualifier="none">Sample Dublin Core Record</dcvalue>
</dublin_core>
```

The `contents` file is a tab-delimited text file listing the files to be included in the item. The first column contains the name of the file, and the second column indicates the bundle into which the file will be placed in the repository. If the file should be displayed in the web interface, place it in the "ORIGINAL" bundle. If there is a specific license file for the item, place it in the "LICENSE" bundle. If you are unsure which bundle to use, use the "ORIGINAL" bundle. Here is an example:

```
contentfile1.pdf    bundle:ORIGINAL
contentfile2.txt    bundle:ORIGINAL
license.txt    bundle:LICENSE
```

The whitespace between the filename and the bundle name must be a single tab character only.

Batch Importing FAQ

How can I check if my XML files are well-formed?

Many XML editors include a tool for checking if a particular document is well-formed. There is also an on-line tool available at the W3C website: http://www.w3.org/2001/03/webdata/xsv. Using this tool upload your `dublin_core.xml` file, and if a "*Low-level XML well-formedness and/or validity processing output*" issue is present then there is a problem to fix. The most common errors are unclosed tags or escape characters. If you need to include any of the following characters inside the text of an element or attribute be sure to use the corresponding escape sequence.

Character	Escaped	Label
"	&quot;	(double) quotation mark
&	&amp;	ampersand
'	&apos;	apostrophe (or single quote)
<	&lt;	less-than sign
>	&gt;	greater-than sign

What character encoding should I use for "dublin_core.xml"?

Unicode, or `UTF-8`, should be your first choice for character encodings. Depending upon what scripting language you are using to create these file you may need to explicitly set the encoding when writing to a file. Any good text editor (such as UltraEdit for Windows or TextMate for OS X) will let you manipulate file encodings, and the '`file`' command on UNIX systems is useful. A second choice for character encoding is "Latin-1" formally known as `ISO-8859-1`. If these are proving difficult and you are on a windows platform, "`windows-1252`" may be another option.

What if I need to import metadata in a schema other than Dublin Core?

First check with the repository administrator to see if another schema is needed. If another schema is needed create a separate file for the other schema named: "`metadata_{prefix}.xml`", where `{prefix}` is replaced with the schema's prefix. Then inside the xml file use the same Dublin Core syntax, but on the `<dublin_core>` element include the attribute "`schema={prefix}`". Here is an example for ETD metadata, which would be in the file "`metadata_etd.xml`":

```
<?xml version="1.0" encoding="UTF-8"?>
<dublin_core schema="etd">
    <dcvalue element="degree" qualifier="department">Computer Science<dcvalue>
    <dcvalue element="degree" qualifier="level">Masters</dcvalue>
    <dcvalue element="degree" qualifier="grantor">Texas A &amp;amp; M</dcvalue>
</dublin_core>
```

What is a bundle and what bundles are available?

Bundles are just groupings of files in the repository, they separate various types of files so that the repository may interact with them specially. There are several standard bundles such as: ORIGINAL, LICENSE, LICENSE_CC, THUMBNAIL, and METADATA. Note that by default, only files contained in the "ORIGINAL" bundle will be listed for download when viewing the item in the repository. Other bundles may be available; check with your repository administrator to know which bundles are appropriate.