Documentation for initial ingest scripts

Here are the expanded comments I wrote from a point of ignorance on Fitz's scripts:

1. html_to_csv.rb
- basically, the FTK html output has information that is needed by multiple steps in the ingest scripts, and putting that info into a csv format is easier for processing that continually parsing it out of the FTK html.

Parses the Forensic Toolkit (FTK) Bookmark html files
and creates a csv file of their aggregation for importing into fedora.
#
These bookmark files have technical metadata, as well as descriptive
metadata that follows a scheme set up in the FTK application by Peter Chan.

2. reorg_directory.rb

Takes the output from html_to_csv.rb (containing the aggregation
of desired data from the FTK bookmark html files), and creates a populated
directory structure easier for FEDORA to ingest.
#
A directory is made for each distinct Fedora object (a file per FTK bookmark
"exportAs" information), containing the relevant FTK output files and the
Transit Solution HTML output files.

3. convert_objects.rb
- this is the meat of the file preparation.

Given the output of reorg_directory.rb, which is a parent directory containing
a (sub)directory for each Fedora object to be created (one per colleciton source file),
this script takes the Transit Solution HTML file within each Fedora object directory,
converts it to a postscript file,
then converts the postscript file into a PDF and multiple per-page jp2000 and text files
for ingest into fedora. These files are output to the appropriate Fedora
object directory with appropriate file extensions.
#
This script requires:
html2ps perl script and the sample profile (http://user.it.uu.se/~jan/html2ps.html)
perl (for html2ps script)
ps2pdf to convert Postscripts to PDF (http://www.ps2pdf.com/)
ghostscript?
ImagicMagick (with jasper + jp2000 libraries installed - http://www.imagemagick.org)
pdftotext (from poppler-utils http://poppler.freedesktop.org/)

Hi everyone,

I had to do some digging on through my backups, but I've resurrected the scripts I used to convert and ingest items into the AIMS application.
I've put the on github over here --> https://github.com/cfitz/aims_scripts

These probably need some explaining....

For the first set of objects that we used (the Gould files), I received a directory of files that Peter had run the FTK toolkit over and had also run a program (I think it was call Avantstar Transit?) that generated HTML files from source files. So, I had to make few scripts that worked with this data in order to get them into fedora.
They are:

html_to_csv.rb : This takes the FTK Bookmark html files that are exported. These bookmark files have technical metadata, as well as descriptive metadata following a scheme that Peter had set up in the FTK application. This script aggrigates all the bookmark HTML files and adds them to a CSV file.

reorg_directory.rb : This script takes the CSV file and makes new directories for each of the objects to be imported into Fedora. It also copies all the source files and Transit converted HTML file for each object into their respective directory.

convert_objects.rb: This script takes the Transit HTML file, converts it to a postscript file, convert the postscript file into a PDF, and multiple per-page jp2000 and text files. This script will require ImagicMagick (with jasper + jp2000 libraries installed), ps2pdf, and perl (for the html2ps script).

aims_ingestor.rb: This script ingest files in the directories into Fedora. Metadata from the CSV file is added into Fedora's XML. Originally, the data model was set so the source object gets it's own fedora object, while the text/jp2000s/pdf/ect files get added as datastreams in a additional "child" fedora object. This script assume that it's being run in the script directory of a rails application that has specific hydra models defined (AimsDocument).

That's about it. I've run these using Ruby REE and they seem to still work fine. The convert objects can take awhile (a few hours), especially if you're using a machine without much processing power.
Let me know if there are any problems or if you have any questions!

best,chris.

FYI - to stanford folks, there's a copy of the files Peter gave me lyberadmin@salt-dev:~/Chris_08_27/ . The file structure I think is an artifact of the Transit application.