Extractors

This page is for the refactoring of the existing extractors. The original wiki page Hosted VMs is still used for the deployments.

As we figure out who's working on what, please start with the following steps for the extractor(s) you chose:

- be able to run the extractor
- add a README, specifically a readme.md (i.e. in markdown), with information on how to install dependencies and run the extractor (in its current shape)
- start looking at dbpedia extractor for template
- Learn about jsonld by playing in the playground here http://json-ld.org/
- Go through the README for the docker extractors template: https://opensource.ncsa.illinois.edu/bitbucket/projects/BD/repos/bd-extractor-template/browse

Steps to take for every extractor in this list:

1. Docker containers
2. JSONLD
3. Extractor info registration
4. Use pyclowder (for python extractors)
5. Add status messages to all extractors and fix level granularity
    a. Make status constants (DONE, ERROR)
    b. Arcgis multiprocessing extractor
6. Register on on demand execution queues
    a. Add on demand key binding to configuration file: messageType = "*.file.text.plain", "extractors."+extractorName
7. Standardize around python logging
    a. Figure out what to log and what format to follow
8. ~~Add logstash to docker compose~~
9. Add sample input/ouput to git repository
10. Add icon for tools catalog to git repository
11. Add entry to Tools catalog, with icon

ID (Extractor Name from config file, same as queue name)	Programming Language	Software	OS	Can be Dockerized?	Assigned To	Repo	Author
DEPLOYED							
ncsa.image.ocr	Python	Tesseract	Linux		Rui	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/ocr	
ncsa.cv.faces	Python	OpenCV	Linux		Rui	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/opencv	Liana
ncsa.cv.eyes	Python	OpenCV	Linux		Rui	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/opencv	Liana
ncsa.cv.closeups	Python	OpenCV	Linux		Rui	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/opencv	Liana
ncsa.cv.profiles	Python	OpenCV	Linux		Rui	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/opencv	Liana
ncsa.cellprofiler.fluorescentcomet	Python	pymedici ?	Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.fly	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.human	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.silvercomet	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.speckle	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.trackobject	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.tumor	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana
ncsa.cellprofiler.yeast	Python		Windows	**No**		https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/cellprofiler	Liana

ncsa.image.sphog	Python	**Matlab**, **mnist-sphog**	Linux		Gregory Jansen	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/handwritten/HandwrittenNumbers	
ncsa.image.caltech101					Gregory Jansen		
ncsa.bisque.histogram (notes: disabled)	Python		Linux				
ncsa.bisque.metadata (notes: disabled)	Python		Linux				
census-section-segmentor	Java		Linux		Sandeep Puthanveetil Satheesan	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/census	Liana, Inna
ncsa.cv.river	Python	**OpenCV** (python), **convert** (from imagemagick), and **Gdal**	Linux		Smruti Padhy	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/river	Liana
ncsa.geo.shpExtractor	Python	gdal	Linux		Jong Lee	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-geo/browse	Jong Lee
ncsa.geo.tiffExtractor	Python	gdal	Linux		Jong Lee	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-geo/browse	Jong Lee
ncsa.image.geotiff	Python	GDAL, Cython, numpy, pygeoprocessing	Linux		Rui	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-geotiff/browse	Rui, Mostafa Elag
ncsa.image.ponddetect	Python	Matlab	Linux		Marcus Slavenas	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-maps/browse/feature_detection	Marcus, Ankit
ncsa.image.humanpref	Python	Matlab	Linux		Marcus Slavenas	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-maps/browse/humanpref	Marcus, Ankit
ncsa.xml.greenindexroute, ncsa.csv.greenindexroute	Python	OpenCV	Linux		Marcus Slavenas	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-maps/browse/greenroute	Marcus
ncsa.image.knn_numerals	Python	OpenCV	Linux		Marcus Slavenas		Marcus
ncsa.audio.speech2text	Java	CMU Sphinx, ffmpeg, sox	Linux		Marcus Slavenas	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-core/browse/audio/speech2text	Marcus
ncsa.audio.preview	Python				Inna	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-core/browse/audio/preview	
ncsa.nlp.simplelanguage	Python	**numpy**			Inna	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-nlp/browse/SimpleLanguage	Liana
ncsa.nlp.simplesummary	Python	**Natural Language Toolkit (NLTK)** for Python, **NLTK Data** or at least: **nltk.corpus**,**nltk.stem.porter** and **nltk.tokenize.punkt**.			Gregory Jansen	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-nlp/browse/SimpleSummary	Liana
ncsa.nlp.SNLPSentiment	Java	Stanford CoreNLP tool, java, maven			Marcus Slavenas	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-nlp/browse/SNLP/SNLPSentimentExtractor	Liana, Marcus(?)
ncsa.nlp.wordtables	Python	**requests**, **pika**, **win32com**				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-nlp/browse/WordTablesExtractor	Liana
siegfried	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-siegfried/browse	Gregory Jansen
ncsa.versus.image	Java	Versus	Linux		Smruti Padhy	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-versus/browse	Kenton, Smruti
ncsa.image.preview (note: check if really deployed. there is an extractor in Hosted VMs list with a similar name.)	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-core/browse/image/preview	Rob, Sandeep
ncsa.pdf.preview (note: check if really deployed. there is an extractor in Hosted VMs list with a similar name.)	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-core/browse/pdf/preview	Rob

ncsa.video.preview (note: check if really deployed. there is an extractor in Hosted VMs list with a similar name.)	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-core/browse/video/preview	Rob
NOT DEPLOYED							
ncsa.image.digitpy	Python	opencv				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/handwritten/SimpleDigitPython	
ncsa.cv.pdfimages		**pdfimages**, from **poppler-utils**				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/poppler	
ncsa.cv.caltech101	Python	**Matlab** and **VLFeat**	64-bit Mac OS			https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cv/browse/vlfeat	
dbpedia	Python	Natural Language Toolkit (NLTK) and rdflib.			Luigi Marini	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-dbpedia/browse	Luigi Marini
digest	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-digest/browse	
ncsa.hpc	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-hpc/browse	Sandeep Puthanveetil Satheesan
LSVA	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-lsva/browse	Liana, Constantinos
LSVA integrated					Sandeep Puthanveetil Satheesan	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-lsva-integrated/browse	Sandeep Puthanveetil Satheesan
ncsa.movieslice	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-movieslice/browse	Sandeep
mri2mesh	Python	pymedici, subprocess, logging, os, numpy, shutil, zipfile				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-mri/browse/mri2mesh	Marcus
msc-ChemCBCExtractor	Python	**requests, pika,** openpyxl, xlrd, pymongo	Linux		Yan Zhao	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-msc/browse/ChemCBCExtractor	Yan
msc-IsletExtractor	Python	**requests, pika,** openpyxl, xlrd, pymongo	Linux		Yan Zhao	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-msc/browse/IsletExtractor	Yan
msc-MonitorExtractor	Python	**requests, pika,** openpyxl, xlrd, pymongo	Linux		Yan Zhao	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-msc/browse/MonitorExtractor	Yan
ncsa.msc.dailymonitor	Python	**requests, pika,** openpyxl, xlrd, pymongo			not used	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-msc/browse/OldMonitorExtractor	Ashwini
msc-PhenotypeExtractor	Python	**requests, pika,** openpyxl, xlrd, pymongo	Linux		Yan Zhao	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-msc/browse/PhenotypeExtractor	Yan
ncsa.nlp.SNLP	Java	Stanford CoreNLP tool, java, maven				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-nlp/browse/SNLP/SNLPExtractor	Liana
ncsa.nlp.tika	Python	Tika project page, pymedici			Kenton McHenry	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-nlp/browse/tika	Liana
person-detector	Python	MATLAB, FFMPEG, requests and pika				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-person-detector/browse/python	Sandeep
ncsa.person-tracker	Python	python, MATLAB, FFMPEG requests and pika				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-person-tracking/browse/python	Sandeep
terra.plantcv	Python	pika requests wheel				https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-plantcv/browse	Yan
medici_PTM_thumbnails	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-ptm/browse/PTMThumbnailExtractor	Constantinos

medici_PTM_metadata	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-ptm/browse/PTMMetadataExtractor	Constantinos
Name not clear **PtmMetadata(?)**	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-ptm/browse/PTMMetadata	Constantinos
medici_ptm_maps	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-ptm/browse/PTMMapsExtractor	Constantinos
medici_ptm_3d	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-ptm/browse/PTM3DExtractor	Constantinos
medici_images_ptm	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-ptm/browse/ImagesPTMExtractor	Constantinos
extractors-rabbitmq **(look like examples)**						https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-rabbitmq/browse	
Name not clear extractors-seabird/	Scala					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-seabird/browse	Luigi
medici_3d_x3d (one of extractors-3d)	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-3d/browse/ObjJSONExtractor	Constantinos
medici_3d_obj_merger (one of extractors-3d)	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-3d/browse/OBJMergerExtractor	Constantinos
medici_oni (one of extractors-3d)	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-3d/browse/OniExtractor	Constantinos
medici_ply_obj (one of extractors-3d)	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-3d/browse/PlyObjExtractor	Constantinos
medici_3d_metadata (one of extractors-3d)	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-3d/browse/ThreeDMetadataExtractor	Constantinos
medici_x3d_html (one of extractors-3d)	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-3d/browse/X3DhtmlExtractor	Constantinos
ncsa.arcgis.landsat7mosaic	Python	ArcGIS	Windows	**No**	Smruti Padhy	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-bd-cz/browse/ndviextractor	Smruti
ncsa.arcgis.floodplain	Python	ArcGIS	Windows	**No**	Smruti Padhy	https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-bd-cz/browse/terex_floodplain/config.py	Smruti
medici_book	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-books/browse/BookPreviewExtractor	Theerasit Issaranon
medici_image_pyramid	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-books/browse/ImagePreviewPyramidExtractor-shebook	Theerasit Issaranon
shebook	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-books/browse/SheBookPreviewExtractor/src/BookPreviewExtractor https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-books/browse/SheBookPreviewExtractor/src/bookpreviewextractor	Theerasit Issaranon
lsva-cedd	Java					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cedd/browse	Constantinos

ncsa.cinemetrics	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-cinemetrics/browse	Constantinos
ncsa.image.metadata	Python					https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/extractors-core/browse/image/metadata	Max. Rob
ncsa.debod.segmentor						https://opensource.ncsa.illinois.edu/bitbucket/projects/DEBOD/repos/extractors-cellsegmentor/browse	
ncsa.image.dmp						https://opensource.ncsa.illinois.edu/bitbucket/projects/DEBOD/repos/extractors-debod/browse https://opensource.ncsa.illinois.edu/bitbucket/projects/DEBOD/repos/extractors-dmp/browse	
ncsa.image.sphog.debod						https://opensource.ncsa.illinois.edu/bitbucket/projects/DEBOD/repos/extractors-handwrittendecimals/browse	
ncsa.image.iarp_remove_circle						https://opensource.ncsa.illinois.edu/bitbucket/projects/IARP/repos/image_fetcher/browse/extractors/remove_circle	Marcus
ncsa.cv.meangrey						https://opensource.ncsa.illinois.edu/bitbucket/projects/IARP/repos/image_fetcher/browse/extractors/mean_grey	Marcus