Generic Newspaper SP

- ***Problem***
    - Need a SP that allows for the batch ingest and transformation of TIFF images that are derived from newspapers. Users will need to browse by newspaper name, year, date, and location. Users will need to keyword search the content. Our current search result model is to provide users a list of newspaper page results. When a page is selected users will be presented with the page in the context of the issue it is from. The search term will be highlighted on the page.

- ***Solution Description***
    - Content Models
        - Newspaper CModel
            - for UPEI leverage the Historical Checklist(HC) document to provide a rich metadata record for each newspaper object.
            - derive an 'essay' for the object ala LoC from the Prospectus section for each newspaper in the HC
                - this could be a MODS Note or Abstract element (e.g. <note type="prospectus"/>)
                - an alternate approach would be to store the essay in a separate datastream. It could be marked up, etc.
            - Datastreams
                - DS-COMPOSITE-MODEL
                - DC [a transformed version of the MODS ds]
                - MODS
                - RELS-EXT
                - TN
                - METSRights ?
                - ESSAY ?
                    - An alternative to the Notes element in the MODS, providing a more powerful approach to describing newspapers with rich text, creating a query to list and display all Essay DSs in a newspaper collection, etc.
                    - In HTML, TEI (with a transform) is an option.
                    - See http://chroniclingamerica.loc.gov/essays/241/
        - Issue CModel
            - Datastreams
                - DS-COMPOSITE-MODEL
                - DC [a transformed version of the MODS ds]
                - MODS
                - RELS-EXT
                - TN
        - Page CModel

    - Datastreams
        - RELS-EXT
        - DC
        - MODS
        - TIFF
            - the archival version of the page image
        - JP2
            - a compressed JP2 that we serve up and derive other image formats at various resolutions
        - TN [thumbnail]
        - OCR
        - HOCR
            - the hOCR as it is generated out of Tesseract provides structural markup for the page image.
            - this could be ABBYYXML or METS/ALTO
            - may also want to specify preferred format in Admin, but has multiple impacts, so may not be this easy
        - ENCODED_OCR
            - a transformed version of the hOCR that we use for term highlighting (eg. lower-cased, …)
        - TECHMD
            - the technical metadata extracted from the TIFF datastream using FITS
- Newspaper Viewer Module
    - Preliminary discussion about newspaper viewer options and Islandora is here: https://docs.google.com/document/d/1eZCU_7PWZy-uuP9zc6euouPv6HjfyGCcZ-kBSoUj5wk/edit
    - Should be a separate drupal viewer module … modeled on the Islandora Viewer framework. Example of a 'viewer' here: https://github.com/Islandora/islandora_jwplayer
    - Main Requirements
        - zoomable
        - works with Djatoka image server
        - supported
        - extendable
        - themeable
- Ingest
    - command line batch
        - UPEI's approach is the use of drush scripts in concert with our server based php listeners.
        - dependent on a .csv of metadata related to the newspapers images being ingested.
    - Add an issue using Islandora's Batch Ingest Module
    - Add an individual page using Add.

        - other ingest methods not yet explored - METS, BagIt, Sword

- ***Use Cases***
    -
    -
- ***Requirements***
    -
- ***Dependencies***
    - Islandora Core, Drush, PHP Listeners/Microservices, Tuque, Solr
    - Djatoka
    - Seadragon
- ***D6/D7***
    - Will be developed in D7.
- ***Issues***
    -
- ***Resource Needs***
    - Dev Team
        - Paul Pound, Lead Dev
        - Donald Moses, PM
        - Kris Bulman, Theming/UI
        - Peter Lux, Fedora/Infrastructure support
        - Melissa Anez, Docs/Use Cases ?
- ***Initial Development***
    - Of potential interest:
        - http://vre.newspapers.upei.ca
- ***Timeline***
    - Content Modelling
        - completed based on UPEI's needs
    - Content Creation/Ingestion
        - currently ingesting newspapers images and anticipate that ingest will be complete by end of December
        - creation of newspaper top level records ongoing
        - Context/storytelling/themes … eg. history through the headlines
    - Viewer
        - SeaDragon Viewer (http://openseadragon.codeplex.com/)
            - Leverage existing work by the Library of Congress.
                - http://sourceforge.net/apps/trac/loc-ndnp/
            - seadragon + djatoka
                - https://github.com/dougreside/DjatokaSeadragon
                - POC - http://vre2.upei.ca/newspaperstest/sites/vre2.upei.ca.newspaperstest/files/seadragon.html
            - 3 weeks of dev/testing
    - Initial Production Test

        - End of November 2012
    - Interface
        - Theme
            - 1 week of dev
        - Search
            - Search term highlight
                - 5 days of dev/testing
        - Browse
            - Calendar based
                - 3.5 days of dev
            - Location based
                - 2 days of dev
        - User Contributed Options
            - Annotations Module (eg. newspaper headlines, sections, etc.; content type - eg. obit, advert, editorial, etc.)
                - 1 week of research
                - 1 week of dev

- ***Interested Parties***
    - DGI
    - University of New Brunswick
    - University of Manitoba
    - Wider Islandora Community