README: Infrastructure Characteristics of re3data.org Research Data Repositories

About

- **Author:** Ceilyn Boyd, ceilyn.boyd@simmons.edu
- **Created:** 3/14/2021
- **Last update:** 3/30/2021

Description

This set of four (4) Python Jupyter Notebooks support the analysis of data about the research data repositories indexed by https://re3data.org to investigate which of Star's (1999) nine infrastructure characteristics do research data repositories exhibit.

0. Directory Hierarchy

The file hierarchy for this dataset is as follows:

- **root**: Toplevel directory
  - **data**: Contains all production data, including XML downloaded from re3data.org and JSON files generated by code in these notebooks
    - **datasets**: Contains JSON datasets generated by code in `functions_re3data.ipynb` that analyze the re3data.org XML metadata
      - `api_dataset.json` : Generated dataset containing data about `r3d:apiType`
      - `consolidated_dataset.json` : Generated dataset containing data about a variety of r3d metadata elements
      - `data_access_dataset.json` : Generated dataset containing data about `r3d:dataAccessType`
      - `data_license_dataset.json` : Generated dataset containing data about `r3d:dataLicense`
      - `data_upload_dataset.json` : Generated dataset containing data about `r3d:dataUploadType`
      - `institution_countries_dataset.json` : Generated dataset containing data about `r3d:institutionCounty`
      - `metadata_standards_dataset.json` : Generated dataset containing data about `r3d:metadataStandard`
      - `one_element_dataset.json` : Generated dataset containing data about r3d metadata elements that take only one value
      - `repository_info_dataset.json` : Generated dataset containing general information about repositories
      - `reopsitory_language_dataset.json` : Generated dataset containing data about `r3d:repositoryLanguage` , the user interface language
      - `repository_type_dataset.json` : Generated dataset containing data about `r3d:type`
      - `software_dataset.json` : Generated dataset containing data about `r3d:software`
    - **json**: Directory for all generated JSON files (each XML metadata file is converted to JSON for processing)
      - `r3d100000001.json`
        - ...
        - `r3d*.xml`
    - **xml**: XML metadata records downloaded from re3data.org via their API
      - `r3d100000001.xml`
        - ...
  - **figures**: Automatically generated figures for different plots
    - `api.jpg` : Plot of `r3d:apiType`
    - `enhanced_publication.jpg` : Plot of `r3d:enhancedPublication`
    - `repository_type.jpg` : Plot of `r3d:type`
    - `software_name.jpg` : Plot of `r3d:software`
    - `start_year.jpg` : Plot of distribution of `r3d:startDate`
    - `versioning.jpg` : Plot of `r3d:verisioning`
  - **tables**: Automatically generated tables of analyses for different metadata elements. These values are used in the text of the study
    - `api_counts.csv` : Table for `r3d:apiType` information
    - `enhanced_publication.csv` : Table for `r3d:enhancedPublication` information
    - `language_counts.csv` : Table for `r3d:repositoryLanguage` information
    - `metadata_standards.csv` : Table for `r3d:metadataStandard` infomation
    - `software.csv` : Table for `r3d:software` information
    - `start_date_intervals.csv` : Table for `r3d:startDate` by intervals information
    - `start_date.csv` : Table for `r3d:startDate` information
    - `subject_counts.csv` : Table for `r3d:subject` information
    - `type_counts_percentages.csv` : Table for `r3d:type` percentage of repositories information
    - `versioning.csv` : Tablel for `r3d:versioning` information
  - **test**: Directory for all test data input (XML) and output (JSON). Mirrors the production directory structure
    - **data**: Diretory for miscellaneous data associated with testing
    - **datasets:** JSON datasets generated by test code; mirror of production `root/datasets directory` . Directory is normally empty, but the following files will appear when tests are run.
      - `api_dataset.json`
      - `consolidated_dataset.json`
      - `data_access_dataset.json`
      - `data_license_dataset.json`
      - `data_upload_dataset.json`
      - `institution_countries_dataset.json`
      - `metadata_standards_dataset.json`
      - `one_element_dataset.json`
      - `repository_info_dataset.json`
      - `reopsitory_language_dataset.json`
      - `repository_type_dataset.json`
      - `software_dataset.json`
    - **json**: Small number of JSON metadata records generated from XML metadata files in the `root/test/xml` directory
      - `r3d100000001.json`
      - ...
      - `r3d*.json`
    - `r3d100000002.xml` : Sample XML file for processing
      - `r3d100000001.xml`
      - ...
      - `r3d*.xml`
    - **xml**: Small number of XML metadata records used for testing
  - `analyze_re3data.ipynb` : Code that analyzes re3data.org XML metadata and produces plots (*.jpg) and tables (*.csv) related to specific metadata elements (e.g. `r3d:software`)
  - `api_re3data.ipynb` : Code to download and process XML metadata using the re3data.org API
  - `functions_re3data.ipynb` : All functions needed to parse re3data.org metadata
  - `README.ipynb` : This file
  - `README.pdf` : PDF of this file
  - `test_re3data.ipynb` : Unit tests

1. API: Download Data from re3data.org

This Jupyter notebook contains the code needed to download data about repositories from re3data.org using their API (https://www.re3data.org/api/doc)

When you want to download a fresh batch of metadata from re3data.org, manually execute the code in this Python Jupyter Notebook.

Caution: Due to the number of metadata files, the download process may take a long time.

2. Functions: Characteristics of re3data.org Research Data Repositories

Python Jupyter Notebook used to process data about the research data repositories indexed by https://re3data.org. For tests of the functions defined here, see the notebook: Unit Tests: Characteristics of re3data.org Research Data Repositories

This Python Jupyter Notebook is loaded automatically by `test_re3data.ipynb` and by `analyze_re3data.ipynb` . There is no need to load this Notebook by itself unless code modifications are made.

To extend the code to support the analysis of new metadata elements, 1) add processing functions to this file (`functions_re3data.ipynb`), 2) add testing functions to the unit test file (`test_re3data.ipynb`), and then 3) add analyses to `analyze_re3data.ipynb` .

3. Analysis: Infrastructure Characteristics of re3data.org Research Data Repositories

This Python Jupyter Notebook is used to analyze data about the research data repositories indexed by https://re3data.org to investigate which of Star's (1999) nine infrastructure characteristics do research data repositories exhibit.

To generate fresh analyses, including plots and tables, run the following Python command:

```
%run "analyze_re3data.ipynb"
```

4. Unit Tests: Characteristics of re3data.org Research Data Repositories

Python Jupyter Notebook containing tests of all functions defined in the notebook: Characteristics of re3data.org Research Data Repositories.

This Notebook loads `functions_re3data.ipynb` directly.

To generate fresh tests, run the following Python command:

```
%run "test_re3data.ipynb"
```

End document.