PPGI PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Introduction to Data Science

MODULE II – PART I

Digital Objects Management

Prof Sergio Serra e Jorge Zavaleta

Background

As you know...

Jupyter Notebooks are composite digital objects

- Used to develop, share, view, and execute interspersed, interlinked, and interactive documentation, equations, visualizations, and code.

… Researchers seeking to deposit reproductible software and data in repositories…

- Expectation → Repositories will provide documentation explaining "what you can deposit", "supported file formats for deposits", "what metadata need to provide", "how to provide this metadata", etc…

- Reality → Expectation is not met by repositories that currently accept software deposits and complex digital objects (e.g. Jupyter Notebooks)

Curatorial practices around Jupyter Notebooks…

Curation and archiving activity needs to be done, not inhibit a future user's need to adapt the code contained within the Notebook file

We will show you…some approaches, techniques and resources that meet researchers' expectations to ensure long-term availability of software in curated archival repositories

1. Deposit Requirements
2. Metadata Requirements
3. Key Curatorial Questions

1- Deposit Requirements

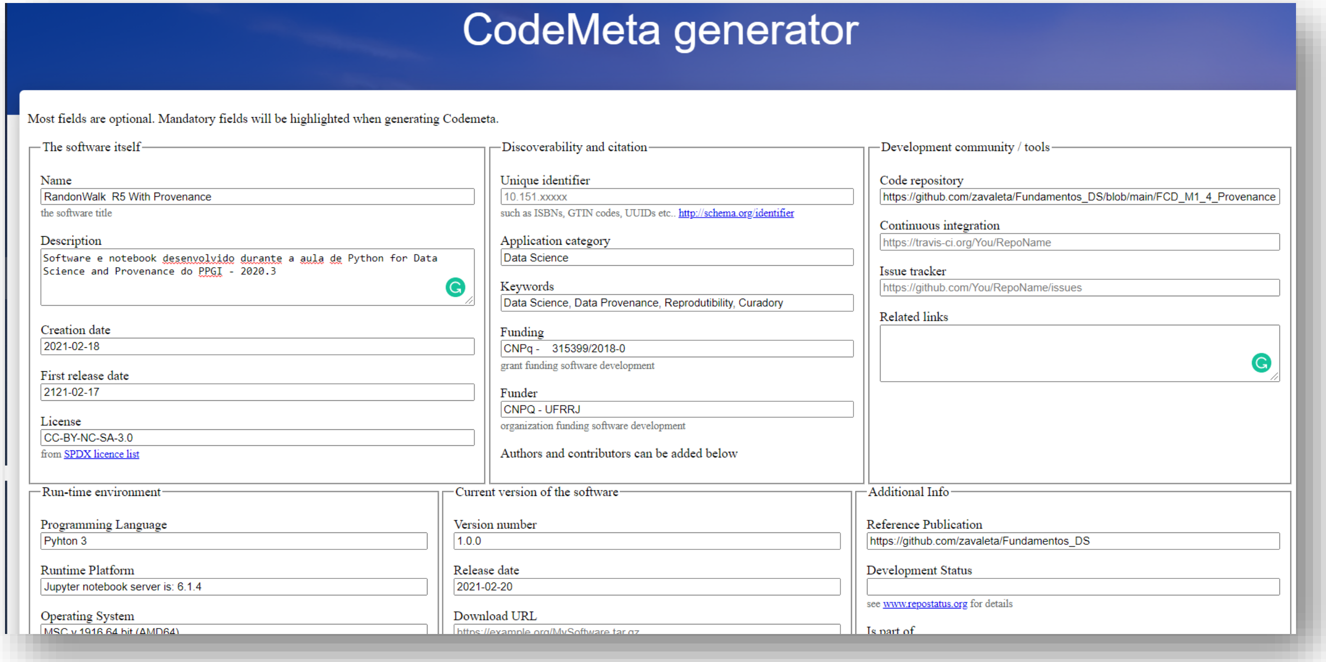
- Minimally required files and metadata will support the ability to open and cite a Jupyter Notebook,
  - Additional functionality is expected without requiring additional files and more comprehensive metadata.

- Minimally required files:
  - .ipynb (cells run with results viewable)
  - README (.txt or .md)
  - LICENSE (.txt or .md)

1- Deposit Requirements (cont)

- Additional functionality is expected without requiring additional files and more comprehensive metadata.
- Additional files to request:
  - PDF of the Jupyter Notebook (export from Jupyter web application) or nbviewer
  - reST (export from Jupyter web application)
  - Sample datasets and provenance (documentation!) (as shown in the last class)
  - CodeMeta.json (https://codemeta.github.io/codemeta-generator/)
  - CITATION.cff (https://citation-file-format.github.io/cff-initializer-javascript/)
  - Container metafile (e.g. docker, singularity, reprozip, binder)
    - Can be published separately with execution instructions; link this to the Jupyter Notebook record
  - Release of the full repository of files associated with .ipynb when applicable and DOI
    - Recommend minting a software DOI for the code repository (e.g. software DOI via Zenodo)

CodeMeta generator



CodeMeta generator

Most fields are optional. Mandatory fields will be highlighted when generating Codemeta.

The software itself

Name: RandonWalk R5 With Provenance
the software title

Description: Software e notebook desenvolvido durante a aula de Python for Data Science and Provenance do PPGI - 2020.3

Creation date: 2021-02-18

First release date: 2121-02-17

License: CC-BY-NC-SA-3.0
from SPDX licence list

Discoverability and citation

Unique identifier: 10.151.xxxxx
such as ISBNs, GTIN codes, UUIDs etc.. http://schema.org/identifier

Application category: Data Science

Keywords: Data Science, Data Provenance, Reprodutibility, Curadory

Funding: CNPq -   315399/2018-0
grant funding software development

Funder: CNPQ - UFRRJ
organization funding software development

Authors and contributors can be added below

Development community / tools

Code repository: https://github.com/zavaleta/Fundamentos_DS/blob/main/FCD_M1_4_Provenance

Continuous integration: https://travis-ci.org/You/RepoName

Issue tracker: https://github.com/You/RepoName/issues

Related links:

Run-time environment

Programming Language: Pyhton 3

Runtime Platform: Jupyter notebook server is: 6.1.4

Operating System: MSC v.1916 64 bit (AMD64)

Current version of the software

Version number: 1.0.0

Release date: 2021-02-20

Download URL: https://example.org/MySoftware.tar.gz

Additional Info

Reference Publication: https://github.com/zavaleta/Fundamentos_DS

Development Status:
see www.repostatus.org for details

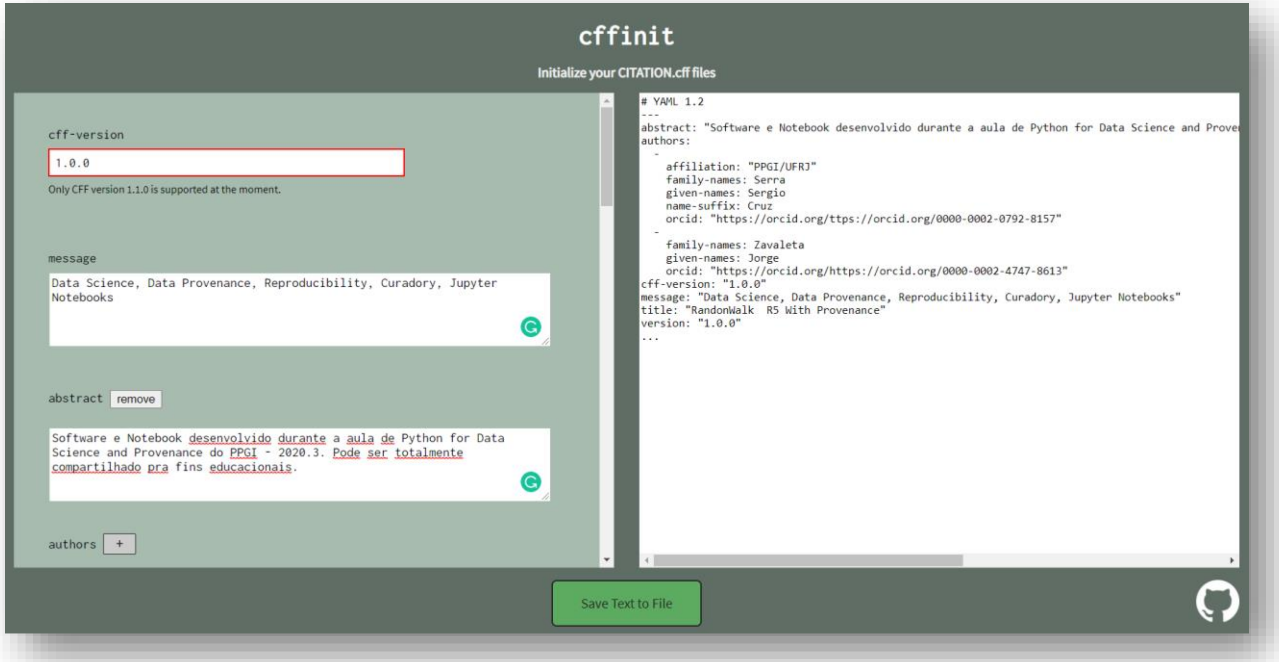
Is part of

https://codemeta.github.io/codemeta-generator/

```
{
   "@context": "https://doi.org/10.5063/schema/codemeta-2.0",
   "@type": "SoftwareSourceCode",
   "license": "https://spdx.org/licenses/CC-BY-NC-SA-3.0",
   "codeRepository":
"https://github.com/zavaleta/Fundamentos_DS/blob/main/FCD_M1_4_Provenance.ipynb",
   "dateCreated": "2021-02-18",
   "datePublished": "2121-02-17",
   "dateModified": "2021-02-20",
   "name": "RandonWalk  R5 With Provenance",
   "version": "1.0.0",
   "description": "Software e notebook desenvolvido durante a aula de Python for Data Science and Provenance do PPGI - 2020.3",
   "applicationCategory": "Data Science",
   "releaseNotes": "Change Log: ALterei nome do app\nBug Fix: Add de comentários",
   "funding": "CNPq -  \t315399/2018-0",
   "isPartOf": "http://www.ppgi.ufrj.br/",
   "referencePublication": "https://github.com/zavaleta/Fundamentos_DS",
   "funder": {
      "@type": "Organization",
      "name": "CNPQ - UFRRJ"
   },
   "keywords": [
      "Data Science",
      "Data Provenance",
      "Reprodutibility",
      "Curadory"
```

CodeMeta.json

Citation.CFF



cff-version

1.0.0

Only CFF version 1.1.0 is supported at the moment.

message

Data Science, Data Provenance, Reproducibility, Curadory, Jupyter Notebooks

abstract remove

Software e Notebook desenvolvido durante a aula de Python for Data Science and Provenance do PPGI - 2020.3. Pode ser totalmente compartilhado pra fins educacionais.

authors +

```
# YAML 1.2
---
abstract: "Software e Notebook desenvolvido durante a aula de Python for Data Science and Prove
authors:
  -
    affiliation: "PPGI/UFRJ"
    family-names: Serra
    given-names: Sergio
    name-suffix: Cruz
    orcid: "https://orcid.org/ttps://orcid.org/0000-0002-0792-8157"
  -
    family-names: Zavaleta
    given-names: Jorge
    orcid: "https://orcid.org/https://orcid.org/0000-0002-4747-8613"
cff-version: "1.0.0"
message: "Data Science, Data Provenance, Reproducibility, Curadory, Jupyter Notebooks"
title: "RandonWalk  R5 With Provenance"
version: "1.0.0"
...
```

```
# YAML 1.2
---
abstract: "Software e Notebook desenvolvido durante a aula de Python for Data Science and Provenance do PPGI - 2020.3. Pode ser totalmente compartilhado pra fins educacionais."
authors:
  -
    affiliation: "PPGI/UFRJ"
    family-names: Serra
    given-names: Sergio
    name-suffix: Cruz
    orcid: "https://orcid.org/ttps://orcid.org/0000-0002-0792-8157"
  -
    family-names: Zavaleta
    given-names: Jorge
    orcid: "https://orcid.org/https://orcid.org/0000-0002-4747-8613"
cff-version: "1.0.0"
message: "Data Science, Data Provenance, Reproducibility, Curadory, Jupyter Notebooks"
title: "RandonWalk  R5 With Provenance"
version: "1.0.0"
...
```

Citation.cff

https://citation-file-format.github.io/cff-initializer-javascript/

2 - Metadata Requirements

Minimal submission: baseline description; enables user to **view and cite the Notebook**

- Jupyter Notebook title
  - Author(s)
  - Jupyter implementation details

    Jupyter version

    Distribution (e.g. Anaconda)

    Kernel version

README

Documents of what the Jupyter Notebook is for!

Request that this file include citation(s) to third-party algorithms and data analyses

Recommend code comments within the Notebook file itself in addition to the README file

License information

2 - Metadata Requirements (cont)

Runnable submission: allows another researcher to execute the Notebook locally using sample data and files provided by the depositor; minimal submission metadata plus:

User documentation

- Instructions to support configuration needed to execute the Notebook and code cells
- Sample input and output files

CodeMeta.json

- Document required software dependencies
- Recommend additional machine actionable dependency documentation (e.g. requirements.txt)

CITATION.cff for the notebook

- Preferred citation; should enable native software citation

Interactive and reproducible repositories powered by Zenodo and Binder.

https://github.com/zavaleta/Fundamentos_DS/blob/main/FCD_M1_5_Zenodo_Binder.ipynb

References

Mendez K. M. (2019) Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. Metabolomics (2019) 15:125 https://doi.org/10.1007/s11306-019-1588-0

Lasser, J. (2020) Creating an executable paper is a journey through Open Science. Communications Physics. https://doi.org/10.1038/s42005-020-00403-4

https://blog.jupyter.org/binder-with-zenodo-af68ed6648a6

https://www.icos-cp.eu/science-and-impact/science-contribution/success-stories/jupyter-notebooks

PPGI PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Introduction to Data Science

MODULE II – PART I

Digital Objects Management

Prof Sergio Serra e Jorge Zavaleta