Zenodo: Preservation meter

August 2014

Author:
João Gonçalves

Supervisor(s):
Lars Holm Nielsen

CERN openlab Summer Student Report 2014

Project Specification

The aim of this openlab summer student project is to enhance ZENODO digital repository service with several preservation-oriented features, such as preservation meter and badge to indicate the suitability of a document for long-term reservation. The project will be developed in the Python programming language, using Flask/HTML5/jQuery/TwitterBootstrap technologies for the user interface and SQLAlchemy/MySQL for persistence.

Abstract

Digital Preservation consists mainly in storing digital information, mostly digital-born content, and making sure that it remains available and accessible in the future. This tasks has many challenges such as making sure that the files are in a known and acessible format, that they are not corrupt, lost or unretrievable.

The digital preservation challenges apply, noticeably, on digital repositories such as Zenodo. Zenodo aims to provide a secure and trusty way of storing data for the long tail of science. This is to say, storing and connecting information that is normaly not available on the main publications, such as the used datasets for a given study or the produced software for a specfic paper.

The goal of this work was to develop a Preservation Meter that allowed the users to know how suitable the files on their submited records are in terms of preservation.

This was accomplished by using a simple and intuitive visual representation of such suitability by means of a progress bar, where a completly filled bar means the file is very likely to be well preserved.

The overall goals of the project were completed and the implementation of this work was integrated on the Zenodo repository as a plugin.

Table of Contents

1 Introduction

Ensuring that valuable digital information remains accessible and usable throughout the years is a topic that is not as trivial as it may seem. There are several initiatives from renowned institutions such as the Library of Congress or the UK Data Archive that, other than just coping with this problem, also provide some best practises and guidelines to interested parties.

The most obvious challenge regarding digital preservation is keeping up with the fast-paced evolution of the hardware and software where digital information is produced and stored. For instance, the standard file format to store a specific kind of information changes as the hardware architectures and operating systems also evolve.

Developed at CERN, Invenio "is a free software suite enabling it's users to run their own digital library or document repository on the web". Some of the instantiations of the Invenio software are the CERN Document Server or the Zenodo Digital Repository. The main objective of the developed project was to enhance the latter with some preservation-oriented features such has having a meter to indicate the suitabilty of a record for long term preservation. The presence of a preservation meter on the records' page should also help to raise awareness of Zenodo's users for the importance of the preservation of the records of their files.

In order to achieve the main objective and calculate a score for each record, there was the need to define some best practices and guidelines that could provide an analytical score to a given record. This was done by gathering information on the most widely used file extensions for a give type of files, and attributing scores to the extensions that were best fit in terms of preservation, eg., were open and not proprietary.

A test-driven-development approach was pursued throughout the whole project, and therefor all of the developed code was tested and validated.

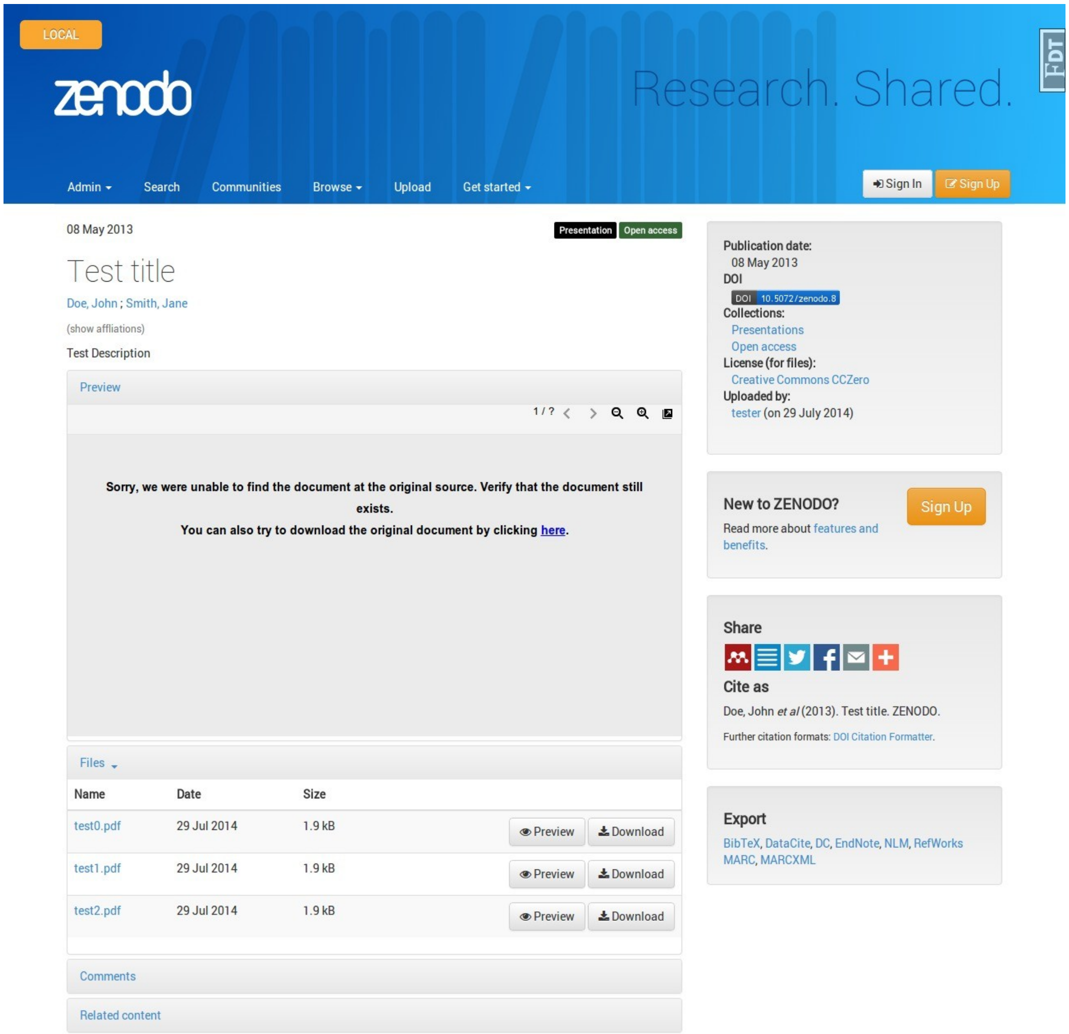
Albeit this report, and the project itself, focuses on the Zenodo Digital repository, all of the work was developed keeping in mind the need to have a plugable and independent module that could be used across multiple instantiations of Invenio, or even on the base Invenio itself.

This report is divided into several chapters, after this introduction, chapter 2 provides a general overview of the developed solution explaining the changes to the existent records page. Chapter 3 provides some detail about architecture and development process of the project. Finally, on chapter 4 some conclusions and future work are discussed.

2 Digital Preservation at Zenodo

This chapter gives an overview of the developed solution, introducing the Zenodo Digital Repository and the enhanced page of the records.

A regular page of a record on Zenodo looks as shown in Figure 1. The record contains some metadata such as title, authors, affiliations and descriptions, as well as a some more technical metadata such as the licence for the files, the DOI of the record or the uploading user.



08 May 2013

Presentation Open access

Test title

Doe, John ; Smith, Jane

(show affliations)

Test Description

Preview

1 / ?

Sorry, we were unable to find the document at the original source. Verify that the document still exists.
You can also try to download the original document by clicking here.

Files

Name	Date	Size	
test0.pdf	29 Jul 2014	1.9 kB	Preview Download
test1.pdf	29 Jul 2014	1.9 kB	Preview Download
test2.pdf	29 Jul 2014	1.9 kB	Preview Download

Comments

Related content

Publication date:
08 May 2013
DOI
DOI 10.5072/zenodo.8
Collections:
Presentations
Open access
License (for files):
Creative Commons CCZero
Uploaded by:
tester (on 29 July 2014)

New to ZENODO?
Read more about features and benefits.

Sign Up

Share

Cite as

Doe, John *et al* (2013). Test title. ZENODO.

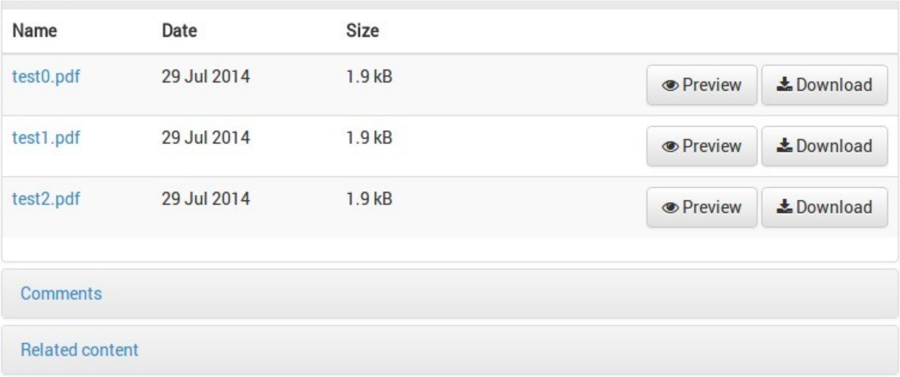
Further citation formats: DOI Citation Formatter.

Export

BibTeX, DataCite, DC, EndNote, NLM, RefWorks MARC, MARCXML

Fig.1: Typical record of Zenodo

However, the most significant part of the record's page is the actual description of it's the included files, as shown in Figure 2.

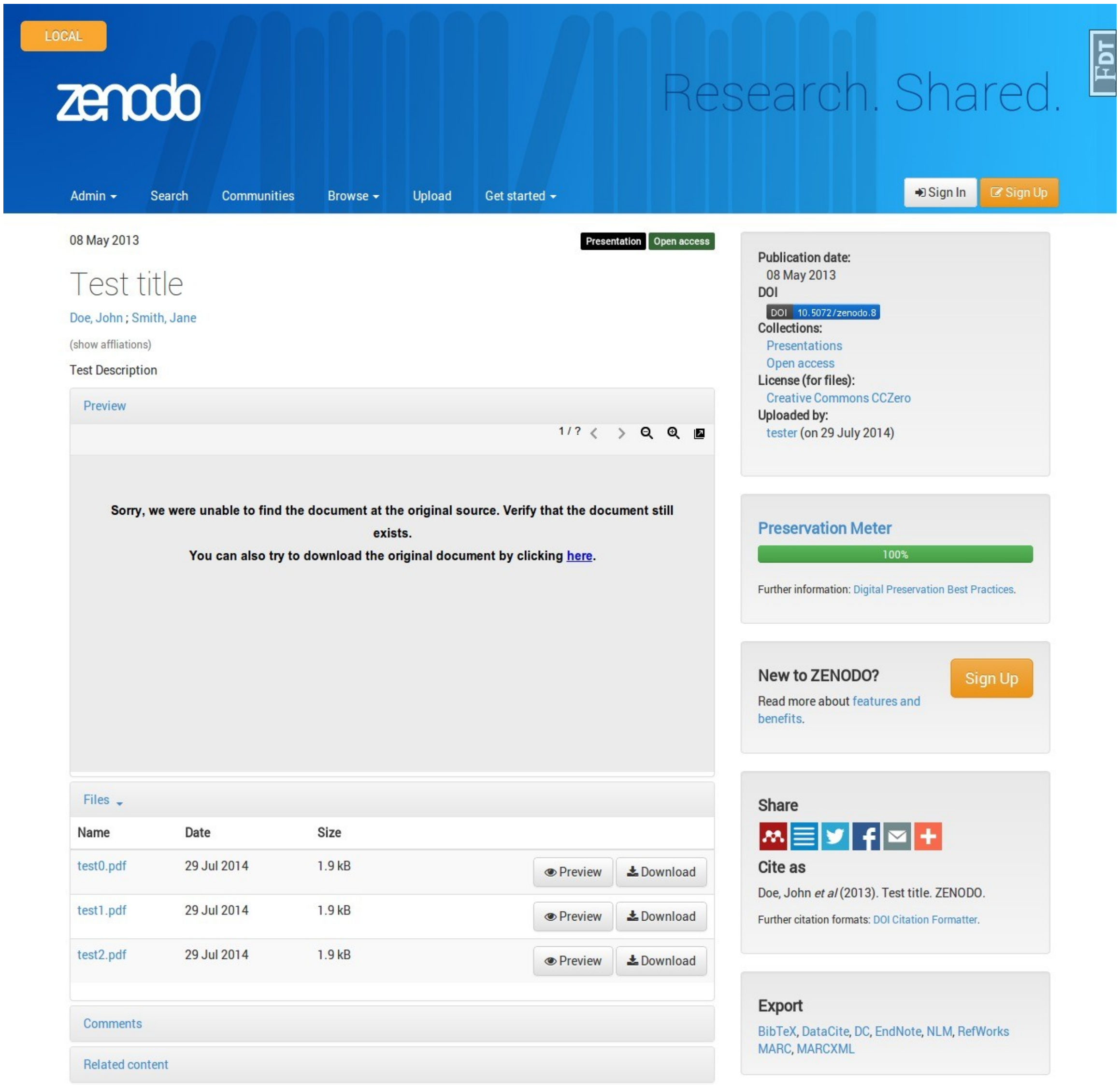


Name	Date	Size
test0.pdf	29 Jul 2014	1.9 kB
test1.pdf	29 Jul 2014	1.9 kB
test2.pdf	29 Jul 2014	1.9 kB

Fig.2: List of files of a record

Based on the files contained on each record, a value for the preservation suitability was calculated based primarily on the extensions of each file.

The new record's page is as described on Figure 3. It includes on the right side a new "well" containing a meter indicating the suitability of the records by means of percentage. It also contains a link to best practices in Preservation and also some details about the record score upon expanding the well.



08 May 2013

Test title

Doe, John ; Smith, Jane

(show affliations)

Test Description

Sorry, we were unable to find the document at the original source. Verify that the document still exists.
You can also try to download the original document by clicking here.

Name	Date	Size
test0.pdf	29 Jul 2014	1.9 kB
test1.pdf	29 Jul 2014	1.9 kB
test2.pdf	29 Jul 2014	1.9 kB

Publication date:
08 May 2013
DOI
DOI 10.5072/zenodo.8
Collections:
Presentations
Open access
License (for files):
Creative Commons CCZero
Uploaded by:
tester (on 29 July 2014)

Preservation Meter
100%
Further information: Digital Preservation Best Practices.

New to ZENODO?
Read more about features and benefits.

Cite as
Doe, John *et al* (2013). Test title. ZENODO.
Further citation formats: DOI Citation Formatter.

Export
BibTeX, DataCite, DC, EndNote, NLM, RefWorks
MARC, MARCXML

Fig. 3: Enhanced record page with meter information

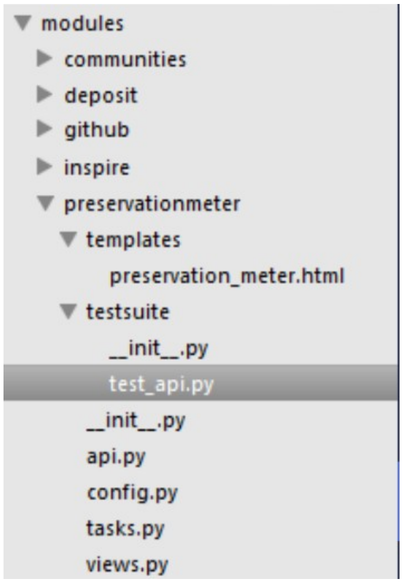
At the moment of the writing of this report, the information about the preservation score is only available after a record is submitted. However, as explained in more detail in the Conclusions chapter, this information will eventually be displayed even before the records' submission.

3 Architecture and Development

This chapters describes in detail the architecture of the developed solution, as well as some development details.

3.1 Architecture

To facilitate the development and deployment of the project, the whole code has been developed as a module for the Zenodo Digital Repository, as shown in an overview of the directory tree of the code in Figure 4.

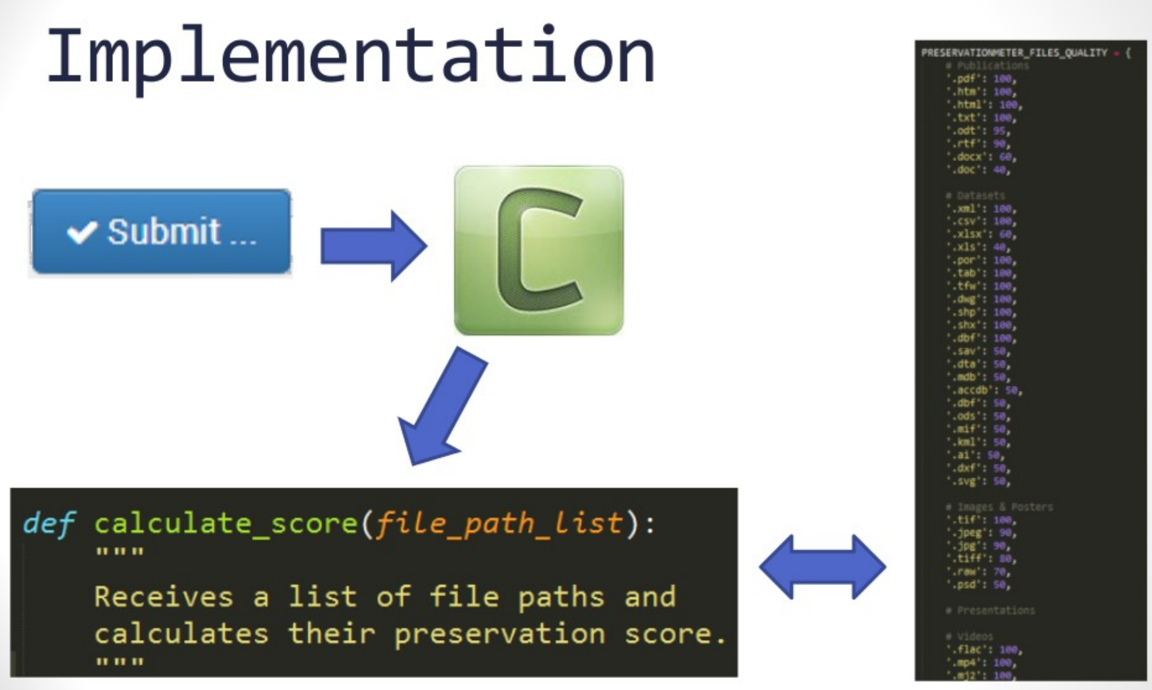


```
▼ modules
  ▶ communities
  ▶ deposit
  ▶ github
  ▶ inspire
  ▼ preservationmeter
    ▼ templates
        preservation_meter.html
    ▼ testsuite
        __init__.py
        test_api.py
      __init__.py
      api.py
      config.py
      tasks.py
      views.py
```

Fig. 4: Directory tree structure of developed module

This approach allows the developed work to be easily plugged in into the most recent version of the Zenodo Digital Repository, while also facilitating the development process in terms of dependency from the latest changes of the related projects. Another important aspect of this approach is the fact that it can also easily adapted for more general purposes, for instance allowing the integration of the module in the Invenio software base code itself, which would in turn allow for a propagation of the module down to the several instances such as the CERN Document Server or the High Energy Physics information system (INSPIRE).

Regarding the actual implementation the score calculation of a record, a small process flow of is depicted on Figure 5.



```
def calculate_score(file_path_list):
    """
    Receives a list of file paths and
    calculates their preservation score.
    """
```

```
PRESERVATIONMETER_FILES_QUALITY = {
    # Publications
    '.pdf': 100,
    '.htm': 100,
    '.html': 100,
    '.txt': 100,
    '.odt': 95,
    '.rtf': 90,
    '.docx': 60,
    '.doc': 40,

    # Datasets
    '.xml': 100,
    '.csv': 100,
    '.xlsx': 60,
    '.xls': 40,
    '.por': 100,
    '.tab': 100,
    '.tfw': 100,
    '.dwg': 100,
    '.shp': 100,
    '.shx': 100,
    '.dbf': 100,
    '.sav': 50,
    '.dta': 50,
    '.mdb': 50,
    '.accdb': 50,
    '.dbf': 50,
    '.ods': 50,
    '.mif': 50,
    '.kml': 50,
    '.ai': 50,
    '.dxf': 50,
    '.svg': 50,

    # Images & Posters
    '.tif': 100,
    '.jpeg': 90,
    '.jpg': 90,
    '.tiff': 80,
    '.raw': 70,
    '.psd': 50,

    # Presentations

    # Videos
    '.flac': 100,
    '.mp4': 100,
    '.mj2': 100,
```

Fig. 5:Process flow of a records' score calculation

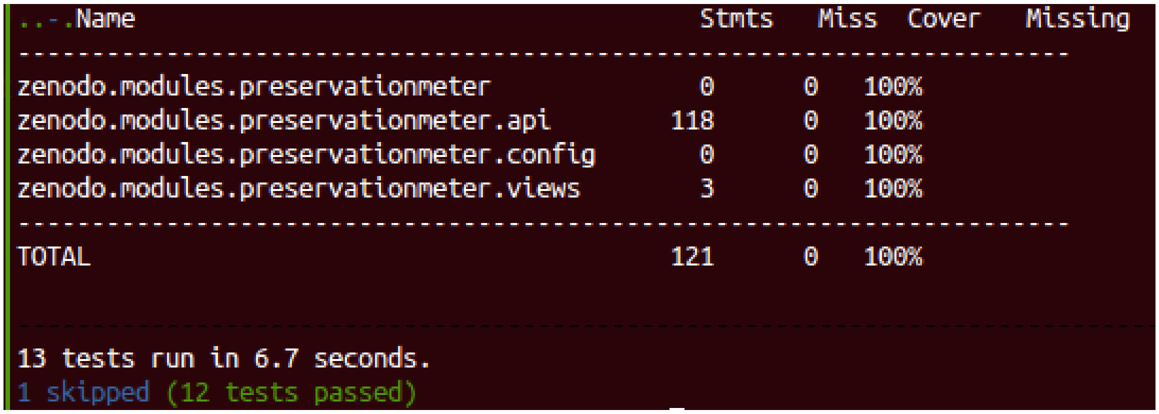
Upon the submission of the record all of the necessary Invenio Celery tasks are triggered. This module also adds an additional Celery task to calculate the score of the record, and therefore the page loading times can be decreased.

The Celery task executes the calculate_score method, passing as input the a list of the paths to the files of the new record. This method then calculates the score of the record by making an average of the score of each file, based on a extension quality map. This calculation could be improved to be something else than an average, as there are some cases where not every file contributes with the same degree of value to the overall record.

3.2 Development

The development of the whole project was done in Linux based operating system. Some software development utilities and quality checkers were used in order to ensure a good workflow and integration with the overall Zenodo Digital Repository.

One of the main aspects of the development was the usage of Test-Driven-Development, which led to good code quality and 100% test coverage, as shown in Figure 6.

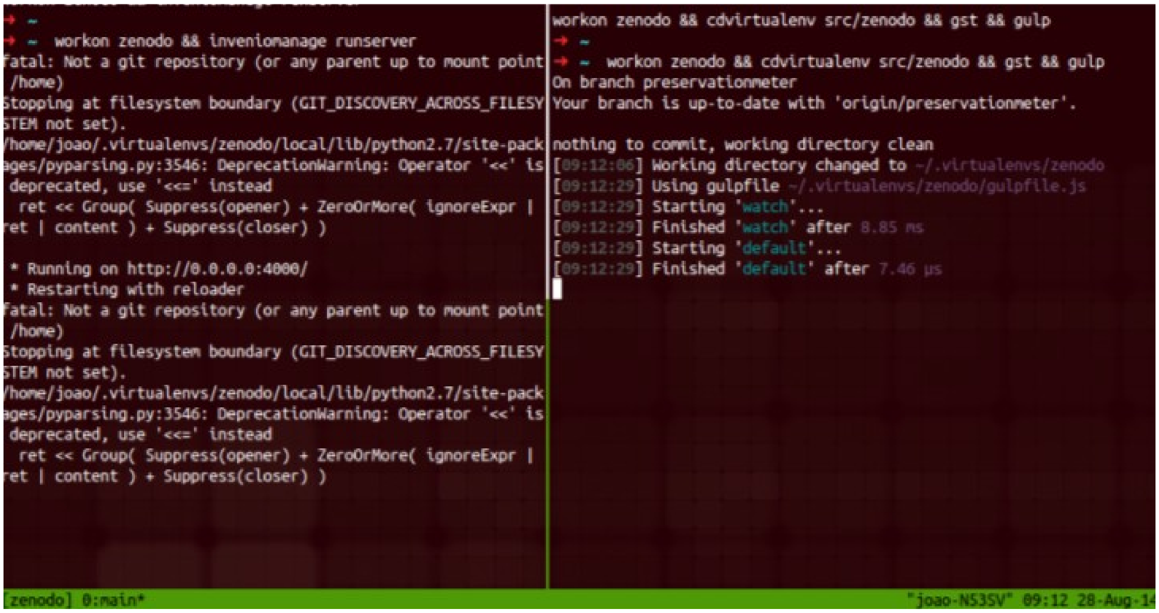


```
..-.Name                                      Stmts   Miss  Cover   Missing
-----------------------------------------------------------------------
zenodo.modules.preservationmeter              0      0   100%
zenodo.modules.preservationmeter.api        118      0   100%
zenodo.modules.preservationmeter.config       0      0   100%
zenodo.modules.preservationmeter.views        3      0   100%
-----------------------------------------------------------------------
TOTAL                                       121      0   100%

-----------------------------------------------------------------------------

13 tests run in 6.7 seconds.
1 skipped (12 tests passed)
```

Fig. 6: Code coverage of the developed project

Each method of the API was tested for both failing and passing scenarios. The testing also resorts to the RecordMocking API in order to be possible to execute them without the need of populating the database with specific records.

Another interesting approach for the development of this project was the usage of the terminal multiplexer, tmux, along with the tmuxinator plugin for automatic session management, as shown in picture 7. This allowed significant improvements in terms of consumed time to start the development and testing.



```
➜ ~
➜ ~  workon zenodo && inveniomanage runserver
fatal: Not a git repository (or any parent up to mount point
 /home)
Stopping at filesystem boundary (GIT_DISCOVERY_ACROSS_FILESY
STEM not set).
/home/joao/.virtualenvs/zenodo/local/lib/python2.7/site-pack
ages/pyparsing.py:3546: DeprecationWarning: Operator '<<' is
 deprecated, use '<<=' instead
  ret << Group( Suppress(opener) + ZeroOrMore( ignoreExpr |
ret | content ) + Suppress(closer) )

 * Running on http://0.0.0.0:4000/
 * Restarting with reloader
fatal: Not a git repository (or any parent up to mount point
 /home)
Stopping at filesystem boundary (GIT_DISCOVERY_ACROSS_FILESY
STEM not set).
/home/joao/.virtualenvs/zenodo/local/lib/python2.7/site-pack
ages/pyparsing.py:3546: DeprecationWarning: Operator '<<' is
 deprecated, use '<<=' instead
  ret << Group( Suppress(opener) + ZeroOrMore( ignoreExpr |
ret | content ) + Suppress(closer) )
```

```
workon zenodo && cdvirtualenv src/zenodo && gst && gulp
➜ ~
➜ ~  workon zenodo && cdvirtualenv src/zenodo && gst && gulp
On branch preservationmeter
Your branch is up-to-date with 'origin/preservationmeter'.

nothing to commit, working directory clean
[09:12:06] Working directory changed to ~/.virtualenvs/zenodo
[09:12:29] Using gulpfile ~/.virtualenvs/zenodo/gulpfile.js
[09:12:29] Starting 'watch'...
[09:12:29] Finished 'watch' after 8.85 ms
[09:12:29] Starting 'default'...
[09:12:29] Finished 'default' after 7.46 μs
```

```
[zenodo] 0:main*                                    "joao-N53SV" 09:12 28-Aug-14
```

Fig. 7 : Automated tmux session management

There was also automation in terms of the unit testing, that were being ran every time a file in a list of watched files, i.e., those being developed, was changed. This automation was achieved with the automated task runner gulp.js.

4 Conclusions

Digital Preservation is an increasingly important and acknowledged challenge. There are several trends and intiatives that aim to address some issues on this task. On this report there was described an approach to allow the users of the Zenodo Digital Repository to be aware of the suitability of their files in terms of long term preservation.

The main objectives of the project, viz., developing a preservation meter to indicate the suitability of a record in terms of long term preservation, while including a best practices guide and developing code testing, were accomplished and the final code was integrated into the Zenodo Digital Repository latest release. The used technologies, such as gulp.js or MockTest allowed for efficient and good quality code development.

As for future and further work, it would be interesting to develop some extra verifications on the file contents itself, as to cross-check the file extension with the MIME-type of the files. Another interesting enhancement would be to include some gamification techniques to help raise awareness of the importance of Digital Preservation and to improve the overall quality of the records. Also important is the improvement of the score calculation in terms of the value of the contribution of each file.