

Организация работы с данными архивов ЦКП «ИКИ-Мониторинг» на основе использования интерактивной среды разработки Jupyter Notebook

Прошин А.А., Бурцев М.А.

Институт космических исследований Российской академии наук, Москва

Современные проблемы дистанционного зондирования Земли из космоса, 13-17 ноября 2023г

Интерактивная среда разработки Jupyter Notebook в последние годы находит все более широкое применение для широкого спектра задач, связанных с анализом и обработкой данных, включая данные дистанционного зондирования Земли. Она, как и следующее ее поколение JupyterLab, реализована в виде удобного веб-интерфейса, позволяющего запускать отдельные участки программного кода и в этом же интерфейсе наглядно представлять результаты операций в виде графиков, изображений или таблиц. Этот инструмент особенно хорошо подходит для решения частных задач, связанных с анализом данных, выработкой алгоритмов, проверкой гипотез и т.п. При этом реализованные в среде документы могут быть наглядным и интерактивным представлением полученных результатов и для других исследователей и разработчиков. Благодаря наличию широкого спектра библиотек, предназначенных для анализа и обработки самых разных типов данных язык программирования Python в настоящее время широко используется не только для прикладных, но и исследовательских задач, в том числе связанных с обработкой спутниковых изображений. Поэтому ниже рассматривается использование именно этого языка программирования в рамках среды Jupyter.

В 2023 году в отделе «Технологии спутникового мониторинга» ИКИ РАН в рамках использования среды Jupyter был реализован базовый функционал для получения доступа к данным сверхбольших распределенных архивов спутниковых данных ЦКП «ИКИ-Мониторинг», позволяющий производить анализ и обработку данных непосредственно по месту их физического хранения. В настоящее время для этого могут быть использованы только различные сторонние библиотеки (gdal, numpy, matplotlib, scipy и др.). Доклад посвящен вопросам организации доступа как к необходимой справочной информации, описывающей имеющиеся в архиве спутниковые данные, так и к самим файлам, содержащим спутниковые наблюдения.

Организация доступа к справочной информации

Для эффективной работы с данными ДЗЗ необходим широкий спектр справочной информации, описывающей спутники, размещенные на них приборы наблюдения и характеристики конкретных информационных продуктов. В рамках разработанной в ИКИ РАН технологии построения сверхбольших распределенных архивов спутниковых данных UNISAT за хранение такой информации отвечает справочная база данных UNISAT_CATALOG. Для доступа к этой информации на языке программирования python была разработана библиотека **uniset_catalog.py**, позволяющая получать различную справочную информацию как в табличном виде, так и в виде структур данных. На скриншоте справа приведены примеры получения справочной информации о продукте, каналах прибора наблюдения, а также о характеристиках каналов конкретных информационных продуктов, имеющих в архивах.

```
[1]: import uniset_catalog as uc
[2]: uc.product(product="rgb", output="table")
```

	product	product_type	product_group	name_en	desc_en	is_metadata
0	rgb	product	seance_syntheses	Surface	Pseudo-color surface image	0

```
[3]: uc.device_bands("MSI", output="table", limit=30)
```

	band_name	type	wl_min	wl_max	resolution
0	1	Coastal aerosol	433.0	453.0	60.0
1	2	Blue	456.0	523.0	10.0
2	3	Green	542.0	577.0	10.0
3	4	Red1	650.0	680.0	10.0
4	5	Red2	698.0	713.0	20.0
5	6	Red3	732.0	747.0	20.0
6	7	Red4	793.0	773.0	20.0
7	8	NIR	785.0	900.0	10.0
8	8A	Narrow NIR1	855.0	875.0	20.0
9	9	Narrow NIR2	935.0	955.0	60.0
10	10	Cirrus	1365.0	1395.0	60.0
11	11	SWIR1	1565.0	1655.0	20.0
12	12	SWIR2	2100.0	2280.0	20.0

```
[4]: uc.device_product_channels(product="rgb", device="MODIS", output="table")
```

	channel_number	datatype	nodata_value	def_rule	calibration_rule	band1_name	band2_name	band3_name
0	1	BYTE	255	{"min":0,"max":255}	{"type":"none"}	7	None	None
1	2	BYTE	255	{"min":0,"max":255}	{"type":"none"}	2	None	None
2	3	BYTE	255	{"min":0,"max":255}	{"type":"none"}	1	None	None

```
[5]: uc.device_product_channels(product="ash", device="MODIS", output="table")
```

	channel_number	datatype	nodata_value	def_rule	calibration_rule	band1_name	band2_name
0	1	BYTE	255	{"min":-3,"max":3,"palette":"ash"}	{"type":"bdt","mult":0.0235294,"add":-3}	31	32

Организация доступа к метаданным спутниковых наблюдений

В рамках технологии UNISAT хранение метаданных, описывающих имеющиеся в архивах экземпляры спутниковых продуктов, реализовано в наборе баз данных unisat, имеющих одинаковую унифицированную структуру. В настоящее время за ведение метаданных архивов ЦКП «ИКИ-Мониторинг» отвечают пять основных баз данных: unisat_hrsat (данные высокого разрешения), unisat_mrsat (среднего разрешения), unisat_comp (композиционные продукты), unisat_radersat (радарные наблюдения), unisat_user (пользовательские продукты).

Справа приведен пример получения метаданных для архива высокого разрешения на основе разработанной для этого библиотеки **unisat_archive.py**. Первый вызов `products_stat` предоставляет статистику по имеющимся в архиве «физическим» продуктам для указанного прибора и диапазона дат. Второй вызов `get_metadata` является основным запросом на получение информации об экземплярах имеющихся в архивах продуктов, соответствующих заданному набору критериев поиска. Результат возвращается в формате JSON и содержит как общую информацию о запросе (секция INFO), так и информацию о найденных экземплярах продуктов (секция DATA). На примере справа также запрашивается информация об относительных путях хранения файлов данных в архиве.

```
[132]: from unisat_archive import Archive
       ua = Archive("hrsat")

[133]: ua.products_stat(device='MSI', from_date='2023-01-01', to_date='2023-11-01', limit=5)

[133]:
```

	product	satellite	station	product_volume	min_date	max_date	count	size_gb
0	channel1	SENTINEL-2A	ESA	ONLINE	2015-07-04	2023-11-02	293329	5077
1	channel2	SENTINEL-2A	ESA	ONLINE	2015-07-04	2023-11-02	293354	177934
2	channel3	SENTINEL-2A	ESA	ONLINE	2015-07-04	2023-11-02	293351	179393
3	channel5	SENTINEL-2A	ESA	ONLINE	2015-07-04	2023-11-02	293323	47517
4	channel7	SENTINEL-2A	ESA	ONLINE	2015-07-04	2023-11-02	293321	48832

```
[134]: ua.get_metadata({
       'bbox': '45.3,45.3,45.5,45.5',
       'dt_from': '2023-07-01',
       'dt': '2023-08-01',
       'devices': 'MSI',
       'satellites': 'SENTINEL-2A',
       'limit': '1',
       'products': 'channel14,channel18',
       'get_online_base_files': '1',
       'volumes': 'ONLINE'
       })

{'DATA': [{'common': {'daynight': 'unknown',
                     'device': 'MSI',
                     'dt': '2023-07-30 07:57:18',
                     'satellite': 'SENTINEL-2A',
                     'station': 'ESA'},
          'products': {'channel14': {'files': ['03737/99718.tif'],
                                       'id': '2307300757185235490110',
                                       'server': 'smisdev'},
                      'channel18': {'files': ['03737/99730.tif'],
                                       'id': '2307300757185235490820',
                                       'server': 'smisdev'}}}],
 'INFO': {'db_pkg_mode': 'hrsat',
          'project': 'smis',
          'query': {'count': 1,
                   'first': 1,
                   'last': 1,
                   'next': 2,
                   'predicted_count': 0,
                   'previous': None,
                   'total_count': 6},
          'servers': {'smisdev': {'accessibility': 1,
                                   'center': 'unknown',
                                   'is_local': 1,
                                   'url': 'dev.nffc.aviales.ru'}},
          'stations': {'ESA': {'tz': '0'}},
          'user': None}}
```

Примеры проведения операций с wybranными экземплярами продуктов

На скриншоте справа приведены примеры операций над файлами данных, соответствующими выбранным на предыдущем слайде экземплярам спутниковых наблюдений. Существенно, что данные примеры являются чисто иллюстративными и предназначены только для демонстрации самых базовых возможностей среды Jupyter.

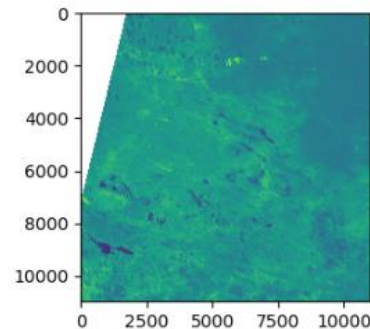
Для получения полных путей к файлам используется получаемый из библиотеки `unSAT_archive` базовый путь к хранилищу файлов `ua.store_dir` и относительные пути, полученные на предыдущем слайде.

В первом примере сначала при помощи библиотеки `gdal` из файлов вычитываются массивы значений для красного и инфракрасных каналов фрагмента изображения по данным прибора MSI спутника Sentinel-2A. Затем в виде одной строки на основе полученных массивов вычисляется вегетационный индекс NDVI. После чего при помощи библиотеки `matplotlib` в блокнот выдается полученное изображение.

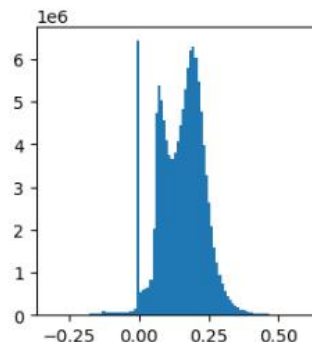
Во втором примере для полученного индекса NDVI также при помощи библиотеки `matplotlib` отображается гистограмма по частоте различных его значений.

```
[147]: # Calculate NDVI
from osgeo import gdal
import matplotlib.pyplot as plt
def read_tif_band1(file):
    dataset = gdal.Open(file, gdal.GA_ReadOnly)
    band = dataset.GetRasterBand(1)
    xval = band.GetNoDataValue()
    arr = band.ReadAsArray()
    res = ma.masked_where(arr == xval, arr)
    return res
red = read_tif_band1(ua.store_dir + '03737/99718.tif')
nir = read_tif_band1(ua.store_dir + '03737/99730.tif')
ndvi = (nir.astype(float)-red.astype(float))/(nir+red)
plt.figure(figsize=(3,3))
plt.imshow(ndvi)
```

```
[147]: <matplotlib.image.AxesImage at 0x7fde2d1e30d0>
```



```
[148]: # Show histogram
plt.figure(figsize=(3,3))
x = plt.hist(ndvi.flatten(), bins=100)
```



Заключение

В настоящее время в отделе «Технологии спутникового мониторинга» реализован только самый базовый функционал, обеспечивающий доступ к данным архивов ЦКП «ИКИ-Мониторинг» и позволяющий работать с ними при помощи широкого спектра сторонних библиотек. При этом доступ к этому инструменту пока что предоставляется только узкому кругу специалистов и разработчиков.

В перспективе планируется радикально расширить имеющийся функционал благодаря поддержке инструментов для анализа и обработки данных, разработанных непосредственно в ИКИ РАН. Существенным шагом в развитии этого инструмента может быть также поддержка механизма «виртуальных» информационных продуктов, которые получаются с использованием заданных правил построения на основе базовых информационных продуктов и являются одним из ключевых преимуществ систем доступа к архивам ЦКП «ИКИ-Мониторинг». Также планируется внедрить специальные программные решения, которые позволили бы предоставить доступ к этому инструменту для более широкого круга пользователей.

Работы по созданию новых инструментов для анализа и обработки данных сверхбольших архивов спутниковой информации с 2019 года выполняются в рамках темы Минобрнауки РФ «Большие данные в космических исследованиях: астрофизика, солнечная система, геосфера» (№122042500019-6) с использованием возможностей ЦКП «ИКИ-Мониторинг» (<http://ckp.geosmis.ru/>).