Grant Agreement no. 640174

PHySIS

Sparse Signal Processing Technologies for HyperSpectral Imaging Systems

INSTRUMENT: Bottom-up space technologies at low TRL
OBJECTIVE: COMPET-06-2014

D8.1 Data management plan

Due Date of Deliverable: 31th August 2015

Completion Date of Deliverable: 6th October 2015

Start date of project: 1st March 2015 Duration: 24 months

Lead partner for deliverable: **FORTH**

Project co-funded by the EC within the Horizon 2020 Programme		
Dissemination Level		
PU	Public	✓
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including Commission Services)	
CO	Confidential, only for members of the consortium (including Commission Services)	

Document History

Issue Date	Version	Changes Made / Reason for this Issue
15 August 2015	V0.1	Initial draft
6 October	V0.2	Final version

Document Main Author(s): Konstantina Fotiadou (FORTH), Konstantinos Karalas (FORTH), Greg Tsagkatakis (FORTH), Bert Geelen (IMEC), Andy Lambrechts (IMEC), Murali Jayapala (IMEC) ,

Document signed off by: Panagiotis Tsakalides (FORTH)

Contents

1. Introduction

1.1. Scope

This document is produced in the framework of the PHySIS project and specifically constitutes deliverable D8.1 within WP8. The scope of this deliverable covers the full extent of the project, since data management is a prevalent subject.

Within this deliverable we consider various low level data management issues including the type of data with will be generated and collected by the project, as well as higher level issues, including the process of data dissemination and preservation. To that end, we consider the guidelines presented by the EC for H2020 projects, as well as capabilities offered by platforms such as OpenAIRE for publication and data storage and public access.

1.2. Purpose

The purpose of this document is to present the methodology for data management, as well as to provide a description of the datasets that have been generated thus far. Regarding the methodology, the document provides an overview of the following issues:

- What types of data will the project generate/collect?
- What standards will be used?
- How will this data be exploited and/or shared/made accessible for verification and reuse?
- How will this data be curated and preserved?

1.3. Applicable documents

[AD 01] PHySIS_Proposal-SEP-210155336

[AD 02] Guidelines on Data Management in Horizon 2020, Version 1.0, 11 December 2013

1.4. Referenced documents

[RD 01] ECSS-E-ST-10C, ECSS Space Engineering – System Engineering general requirements – (issued on 6 March 2009)

[RD 02] ECSS-E-ST-40-06C, ECSS Space Engineering – Software – (issued on 6 March 2009)

[RD 03] CCSDS 311.0-M-1, Reference Architecture For Space Data Systems, September 2008

[RD 04] Space Mission Analysis and Design, James R. Wertz and Wiley J. Larson, eds., Microcosm Press/Springer,1999 (Third Edition)

[RD 05] Hyperspectral Remote Sensing – Principles and Applications M.Borengasser, W.S. Hungate, R. Watkins, CRC Press 2008

[RD 06] Canadian Hyperspectral Spaceborne Mission – Applications and User Requirements K.Staenz, A.Hollinger, 3rd EARSeL Workshop on Imaging Spectroscopy, Herrsching, 13-16 May 2003

[RD 07] Oscar Carrasco, Richard Gomez, Arun Chainani, Willian Roper: Hyperspectral Imaging Applied to Medical Diagnoses and Food Safety. In: Proceedings of SPIE - The International Society for Optical Engineering (Impact Factor: 0.2).08/2003; DOI: 10.1117/12.502589.

[RD 08] D. G. Ferris et al.: Multimodal hyperspectral imaging for the noninvasive diagnosis of cervical Neoplasia, In Journal of Lower Genital Tract Disease. Vol. 5(2), 65–72 (2001).

[RD 09] Robert Koprowski, Sławomir Wilczyński, Zygmunt Wróbel, Sławomir Kasperczyk and Barbara Błońska-Fajfrowska: Automatic method for the dermatological diagnosis of selected hand skin features in hyperspectral imaging. In: BioMedical Engineering OnLine 2014.

[RD 10] Dmitry Yudovsky, Aksone Nouvong, Laurent Pilon: Hyperspectral Imaging in Diabetic Foot Wound Care. In: Journal of Diabetes Science and Technology, Volume 4, Issue 5, September 2010.

[RD 11] Hamed Akbaria, Luma V. Haliga, Hongzheng Zhangb, Dongsheng Wangb, Zhuo Georgia Chenb, and Baowei Feia: Detection of Cancer Metastasis Using a Novel Macroscopic Hyperspectral Method. In: Proc SPIE. 2012; 8317: 831711. doi:10.1117/12.912026.

[RD 12] Guolan Lu, Luma Halig, Dongsheng Wang, Zhuo Georgia Chen, and Baowei Fei: Spectral-Spatial Classification Using Tensor Modeling for Cancer Detection with Hyperspectral Imaging. In: Proc SPIE. 2014 March 21; 9034: 903413–. doi:10.1117/12.2043796.

1.5. Definitions, acronyms and abbreviations

EO-1:	Earth Observation 1
ESA:	European Space Agency
HSI:	Hyperspectral Imaging
HYP:	Hyperspectral
PHySIS:	Sparse Signal Processing Technologies for HyperSpectral Imaging Systems
TBC:	To Be Confirmed
TBD:	To Be Defined
VNIR:	Visible and Near InfraRed
USGS:	United States Geological Survey

2. Data management plan description

The data management plan concerns the datasets generated by the project with respect to four key attributes: i) a description of the datasets; ii) a description of the standards and metadata associated with these datasets; iii) the method that will be employed for sharing these datasets; and iv) a plan for the long term archiving of these data. More specifically:

The acquired or produced data will be associated with a description of the type of data and possible uses of such data. More specifically, the description must include information regarding the spectral range associated with each encoded spectral band, the type of object that is being imaged, the conditions of imaging, i.e., indoors or outdoors, and additional information along these lines. Furthermore, ideally some brief description of possible uses of these datasets will also be included.

In addition to the high level description, the dataset will also include additional information regarding the standards that were followed during acquisition, as well as additional metadata. One type of metadata that is considered in PHySIS is programming code, such as Matlab scripts, while information regarding the low-level processing of the raw data will also be included.

For sharing the datasets internally, the PHySIS consortium will rely on the FTP-based filestore that has been created by the FORTH team. The filestore allows for secure access to the datasets only to the consortium members, aiming at promoting the better understanding and the exploitation of methods developed by each partner.

Long term archiving of the acquired datasets is very important both in terms of visibility past the end of the project, as well as for a greater proliferation in the research community. Within PHySIS, we will consider various archiving platforms, including the PHySIS website[1], the EU OpenAIRE[2] platform, and the UCI machine learning repository[3].

[1] http://www.physis-project.eu/

[2] https://www.openaire.eu/

[3] https://archive.ics.uci.edu/ml/datasets.html

3. PHySIS Datasets

3.1. Satellite based Land Cover Dataset

3.1.1. Description

NASA's MODIS Earth Observation System is considered one of the most valuable sources of remote sensing data, aimed at monitoring and predicting environmental dynamics. The MODIS sensor can achieve global coverage with high temporal resolution, since it is able to scan the Earth's surface (aboard the Terra and Aqua satellites) with a 16-day cover cycle. As far as spectral resolution is concerned, MODIS acquires data in 36 spectral bands ranging from 400 to 14400nm. Note that the first two bands (600 - 900nm) have a spatial resolution (pixel size at nadir) of 250m, bands 3 - 7 (400 - 2100nm) of 500m, and all the rest (400 - 14400nm) of 1km. MODIS data are open-access and continuously updated since 2000.

The CLC inventory was initiated in 1990 and has been updated in 2000 and 2006, while the latest version of the 2012 update is still under production. CLC consists of 44 classes, including artificial surfaces, agricultural and forest areas, wetlands, and water bodies. In this work, we utilize data from 2000 and 2006 at 100m2 resolution (Version 172).

3.1.2. Standards & Metadata

MODIS native product files come in Hierarchical Data Format (HDF) and in SINusoidal (SIN) projection. As a result, MODIS data are grouped in 460 equal non-overlapping spatial tiles starting at (0,0) in the upper left corner and proceeding to the right (horizontal) and downward (vertical) until the lower right corner at (35,17). Each one of them captures approximately 1200x1200 km of real land. Nevertheless, SIN projection is not widely used and thus a common geographic projection is needed for our study. For this reason, we utilized the MODIS Reprojection Tool [1] (MRT1), which provides a basic set of routines for transformation of MODIS imagery into standard geographic projections. This way, we re-sampled the original data and changed the projection to Universal Transverse Mercator (UTM) to become compatible with the global coordinate system (WGS84 datum), which is also used by the Global Positioning System (GPS). The area of interest is comprised of a central portion of the European continent, namely h19v04 (with the exception of regions from Ukraine and Moldova) and h18v04 image tiles (see Fig. 1).

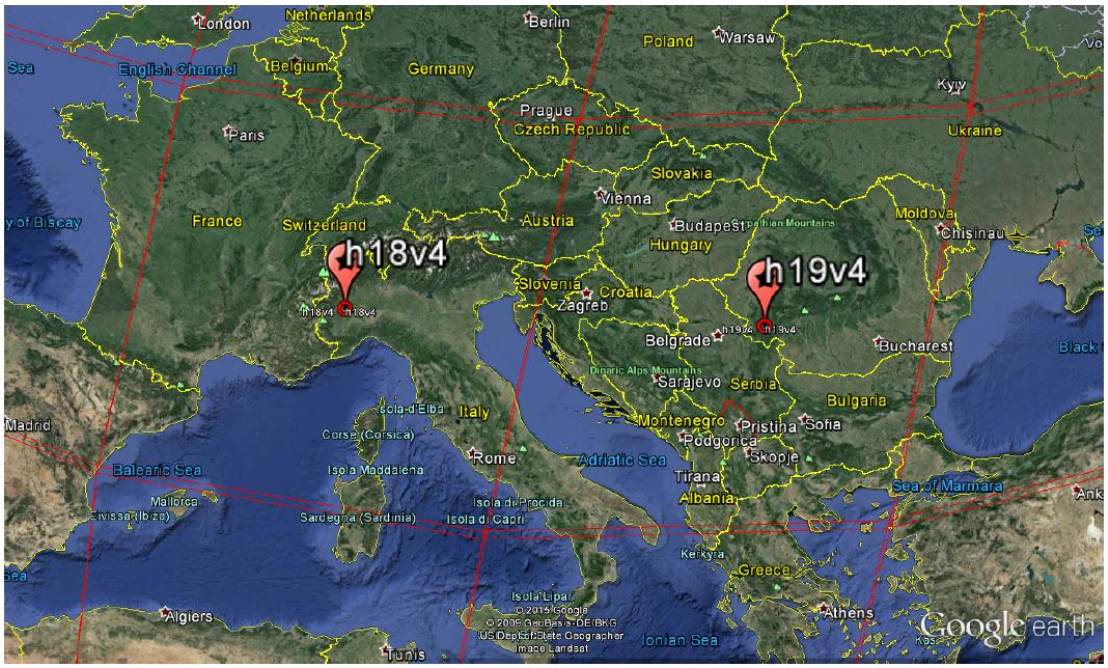


Figure 1: Geographic distribution of MODIS h18v04 and h19v04 tiles. The h18v04 region captures South-Central Europe, while h19v04 a large part of the Balkans. These exemplary regions were selected due to the diversity of land cover and the availability of data.

In order to benefit from the high temporal resolution observations of MODIS and monitor the best possible density and intensity of green vegetation growth, while mitigating lower spatial resolution, we consider annual time series. Our model takes into account the Normalized Difference Vegetation Index (NDVI) from the Level-3 product MOD13A1 Collection 5 (500m, 16 days temporal granularity). NDVI is often used as a monitoring tool for vegetation health and dynamics [2], because it is more sensitive than a single wavelength, as it is calculated from reflectance measurements in the red and near infrared (NIR) portion of the spectrum. In addition to NDVI, Land Surface Temperature (LST) has also been proven to play a significant role in detecting several climatic, hydrological, ecological and biogeochemical changes [8], which are crucial parameters for land cover [9]. We enhance the previously selected examples by adding features related to the LST daytime, extending the number of features (obtained by the MVC technique) to 20. The temperature data are included in the 1km MOD11A2 product (Level-3). In order to obtain the same spatial resolution with MOD13A1, an oversampling to 500m spatial resolution is performed.

For the CLC data, The QGIS3 software is employed in order to transform these raster-based Geographic Information System (GIS) data [10] to WGS84 format, in order to become compatible with MODIS data, and subsequently extract the regions corresponding to the h19v04 and the h18v04 tiles through upper left and lower right latitude and longitude coordinates. In order to construct the multi-label dataset, the CLC labels matrix was divided into non-overlapping blocks using a 55 grid, since the MODIS pixel size is approximately 25 times the size of a Corine pixel as shown in Fig. 2. As a result, a binary vector per sample is produced, where a value of one indicates that a

label is present while a value of zero denotes that a label is absent. We select 20 labels as depicted in Table 1 and exclude examples composed of only one label in order to acquire a challenging scenario for the multi-label learning algorithms.

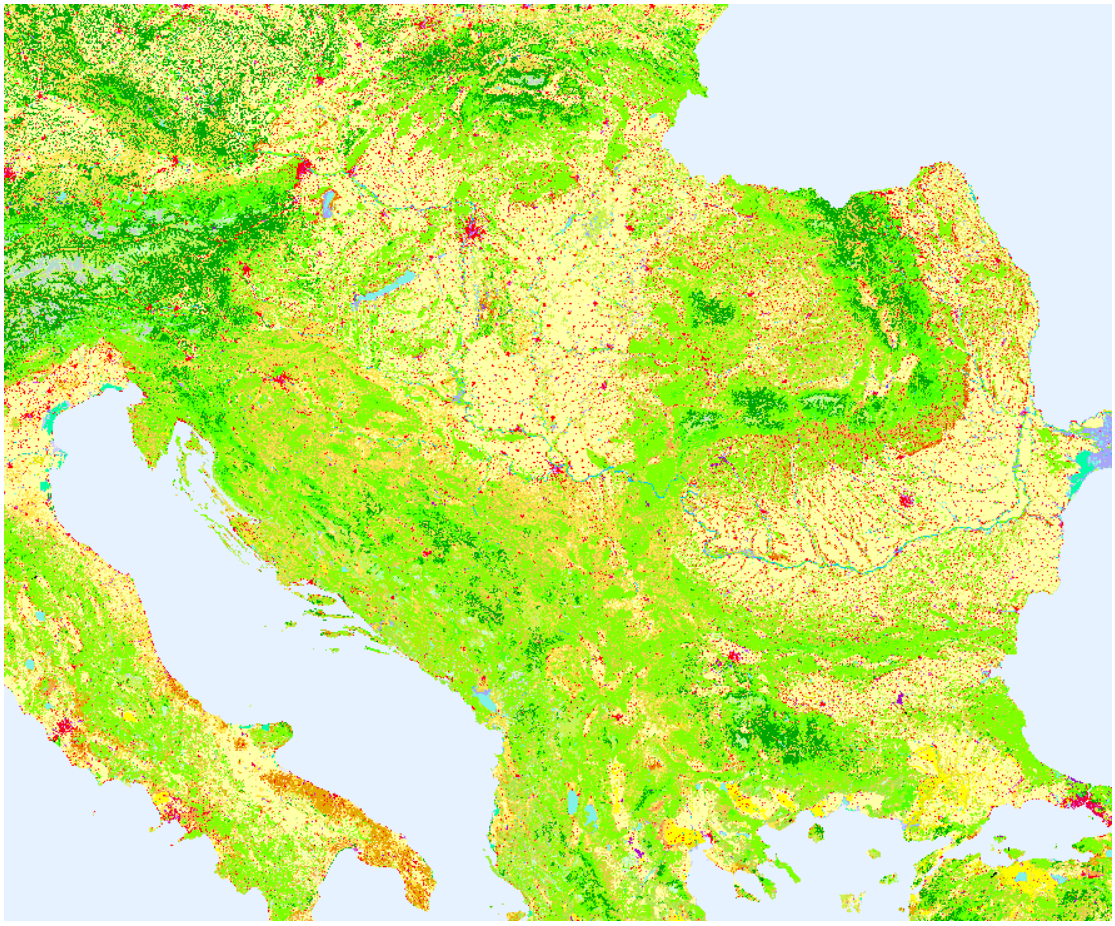


Figure 2: Corine Land Cover map for h19v04 of 2000

Table 1: Ground truth labels

No	CLC code	Description
1	111	Continuous urban fabric
2	121	Industrial or commercial units
3	122	Road and rail network
4	124	Airport
5	131	Mineral extraction sites
6	132	Dump sites

7	133	Construction sites
8	141	Green urban sites
9	142	Sports and leisure facilities
10	212	Permanently irrigated land
11	213	Rice fields
12	223	Olive groves
13	241	Annual crop
14	322	Moors
15	331	Beach, sand
16	332	Bare rocks
17	411	Inland marshes
18	412	Peat bogs
19	421	Salt marshes
20	521	Coastal lagoons

3.1.3. Data Sharing

Data sharing will take place using the PHySIS filestore to the members of the consortium. The PHySIS filestore employs a password protection in order to limit access to unauthorized personnel. The data are encoded in Matlab based files (*.mat files) that can be loaded using the associated Matlab script that will be included.

Public dissemination will be achieved through the PHySIS public website, as well as the institutional websites from consortium members including SPL-FORTH. In this scenario, only a limited part of the dataset will be freely available, under explicitly stated legal conditions concerting their use and the necessary citations.

3.1.4. Archiving

Long term archiving will be sought using publicly available platforms, including the SPL-FORTH website, the UCI machine learning repository and the OpenAIRE platform. The utilization of these platforms will support the long lasting dissemination of the dataset,

free of charge. In general, the limited datasets that will be considered for long-term storage will be in the order of a few Gigabytes or less, thus removing the need for specialized solutions.

3.2. Simulated Satellite Snapshot Mosaic Data based on Hyperion Data

3.2.1. Description

The objective of the Simulated Snapshot Mosaic Data is to utilize currently available multispectral satellite imagery in order to generate artificial data that approximate the imagery that would be acquired by a HYP-enabled satellite equipped with the IMEC spectral mosaic camera. To that end, imagery from the EO-1 satellite, and more specifically, spectral imagery captured by the Hyperion sensor were selected and modified appropriately.

3.2.2. Standards & Metadata

The Hyperion imaging sensor is one of the three principal instruments aboard the EO-1 mission, as part of the National Aeronautics and Space Administration (NASA) New Millennium Program (NMP), acquiring spectral information in both the visible to near infrared (VNIR) and the short wave infrared (SWIR) regions through two spectrometers and a single telescope. [1, 2, 3] Hyperion provides earth observation imagery at 30m spatial resolution, with 7.5 km swath width, in 220 unique spectral channels at 10nm spectral resolution. The sensor provides a full spectrum from the 355.589 to 2577.07 nm. The first 70 spectral channels lie in the VNIR region, while the rest 172 into the SWIR region. Figure 3 illustrates a typical example of Hyperion 29, 23 and 16 spectral bands displaying the Hawaii Island acquired on the 8-30-2015.

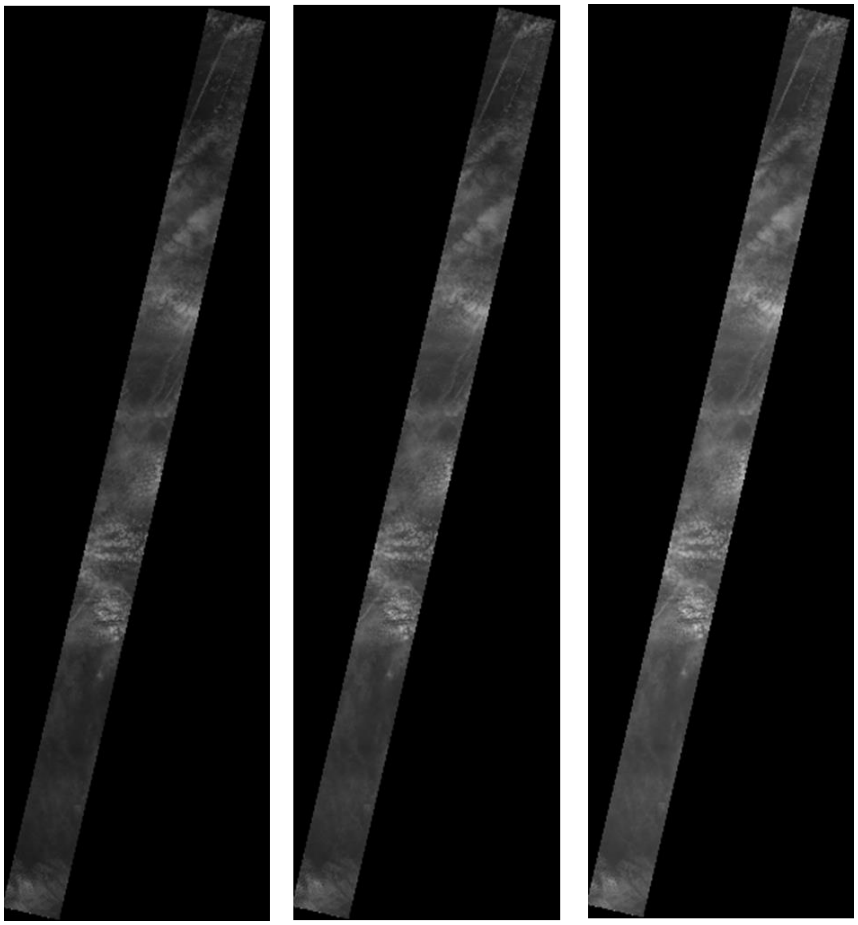


Figure 3: Gray Scale display example of Hyperion data, Hawaii Island: 8-30-2015.

The Level 1 radiometric product includes 242 spectral bands from which only 198 are calibrated. The non-calibration phenomenon is basically caused by the detector's low responsivity, while the non-calibrated channels are simply set to zero. Additionally, due to an overlap among the VNIR and the SWIR focal planes, only 196 spectral bands are unique. Calibrated spectral bands range from the 8-57 for the VNIR, and 77-224 for the SWIR regions. Table 1 illustrates the spectral coverage of the Hyperion sensor.

In Table 1, we provide the mapping between the different spectral bands acquired by the Hyperion sensor and the corresponding bands captured by the IMEC mosaic sensor. Due to different design characteristics, however, some of the bands captured by the IMEC sensor do not map exactly to the bands of Hyperion. To address this issue, the table also provides the relative weighting that is employed in order to approximate the entire spectral cube. Furthermore, data collected by Hyperion have a radiometric resolution of 16bits/sample, whereas the IMEC camera operates at 8bit/sample.

Table 2: Correspondence between IMEC mosaic and Hyperion spectral Bands

Imec Mosaic Camera wavelength	Weighting	Hyperion Band	Average Wavelength (nm)	Full Width at Half the Maximum FWHM (nm)	Spatial Resolution (m)
604.31	0.5	B25	599.8000	10.5607	30
	0.5	B26	609.9700	10.4823	30
617.05	1	B27	620.1500	10.4147	30
625.56	0.5	B27	620.1500	10.4147	30
	0.5	B28	630.3200	10.3595	30
633.72	1	B28	630.3200	10.3595	30
642.86	1	B29	640.5000	10.3188	30
650.51	1	B30	650.6700	10.2942	30
659.74	1	B31	660.8500	10.2856	30
663.56	0.7	B31	660.8500	10.2856	30
	0.3	B32	671.0200	10.2980	30
667.84	0.4	B31	660.8500	10.2856	30
	0.6	B32	671.0200	10.2980	30
678.28	0.4	B32	671.0200	10.2980	30
	0.6	B33	681.2000	10.3349	30
702.78	1	B35	701.5500	10.4592	30
717.34	0.5	B36	711.7200	10.5322	30
	0.5	B37	721.9000	10.6004	30
729.52	1	B38	732.0700	10.6562	30
743.55	1	B39	742.2500	10.6933	30
755.92	1	B40	752.4300	10.7058	30
769.35	0.3	B41	762.6000	10.7276	30
	0.7	B42	772.7800	10.7907	30
781.15	1	B43	782.9500	10.8833	30
794.00	1	B44	793.1300	10.9938	30

812.96	1	0B46	813.4800	11.1980	30
825.15	1	B47	823.6500	11.2600	30
835.37	1	B48	833.8300	11.2824	30
846.43	1	B49	844.0000	11.2822	30
856.30	1	B50	854.1800	11.2816	30
866.35	1	B51	864.3500	11.2809	30
870.60	0.4	B51	864.3500	11.2809	30
	0.6	B52	874.5300	11.2797	30

The USGS earth explorer is a portal that allows the acquisition of a large number of satellite imagery from different time locations, time instances, and instruments. Figure 4 presents an example of selecting data from Hyperion over a region of Western Crete, acquired during March 2005. We utilized these data in order to generate the simulated imagery.

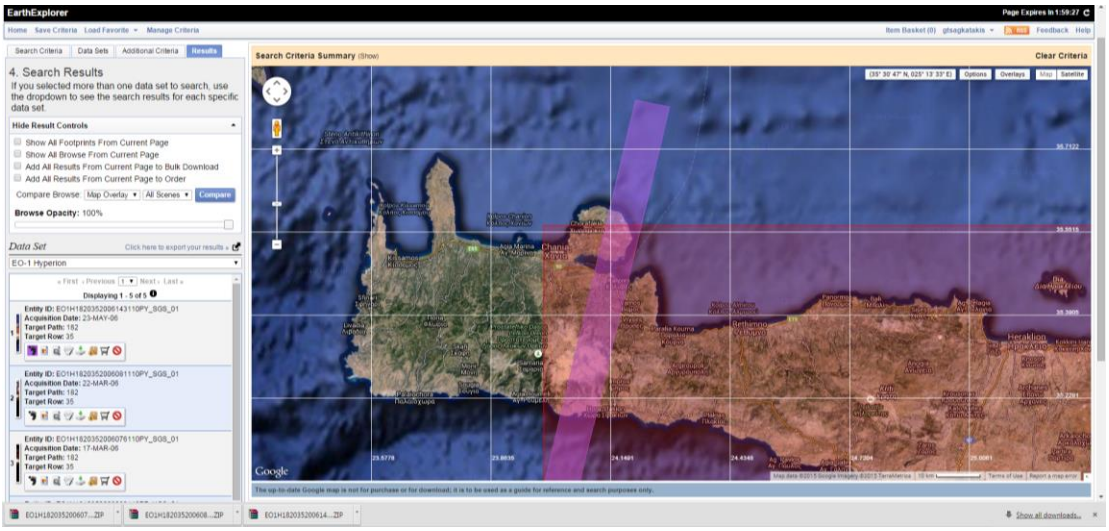


Figure 4: USGS EarthExplore interface and data selection

Figure 5 presents some example imagery from the simulated full-resolution spectral cube that would be generated by a camera sampling at the same spectral bands as the IMEC Mosaic Camera. However, to truly produce a realistic spectral mosaic frame that the IMEC camera is able to produce, we must enforce the spectral mosaic pattern.

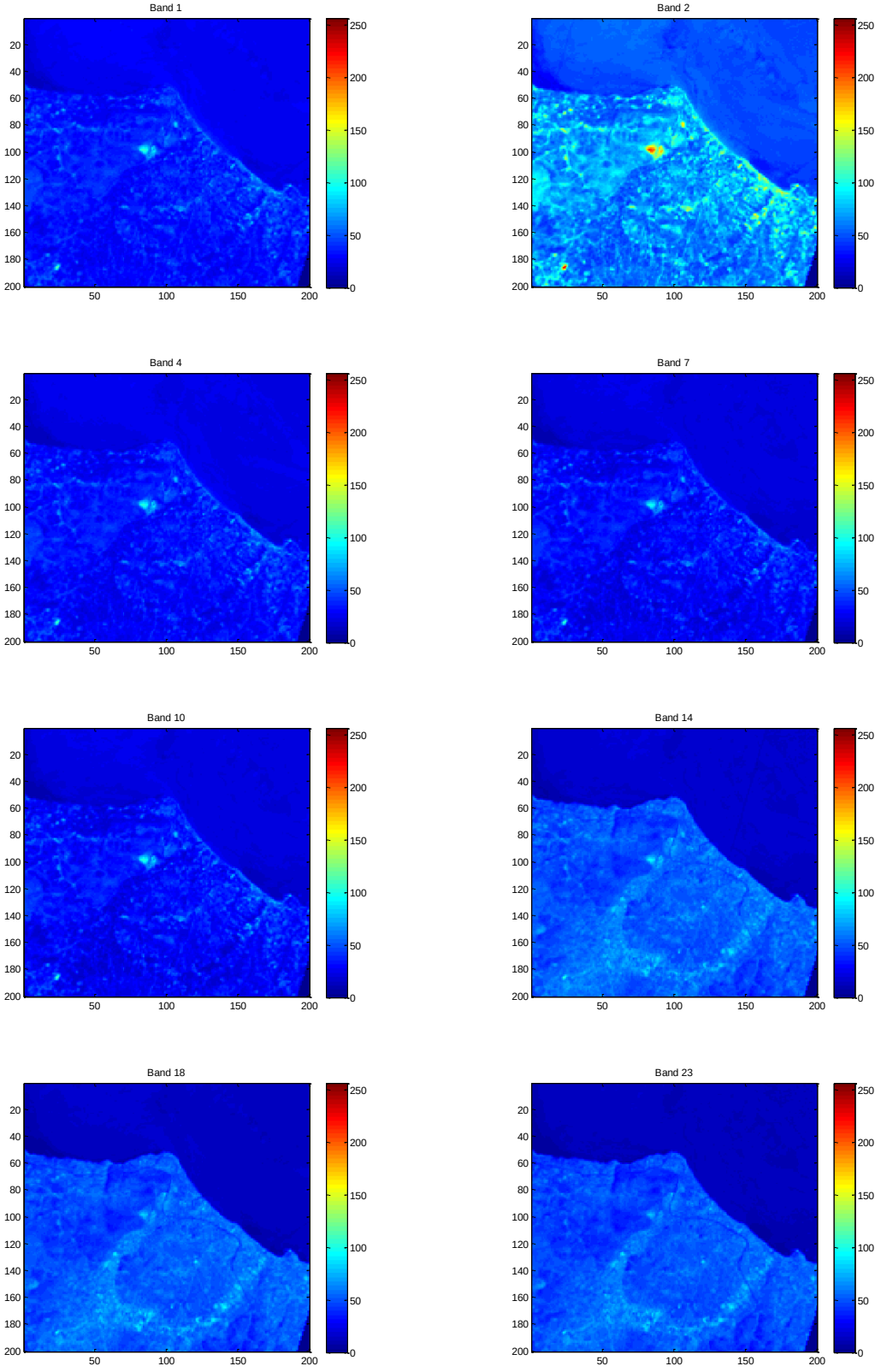
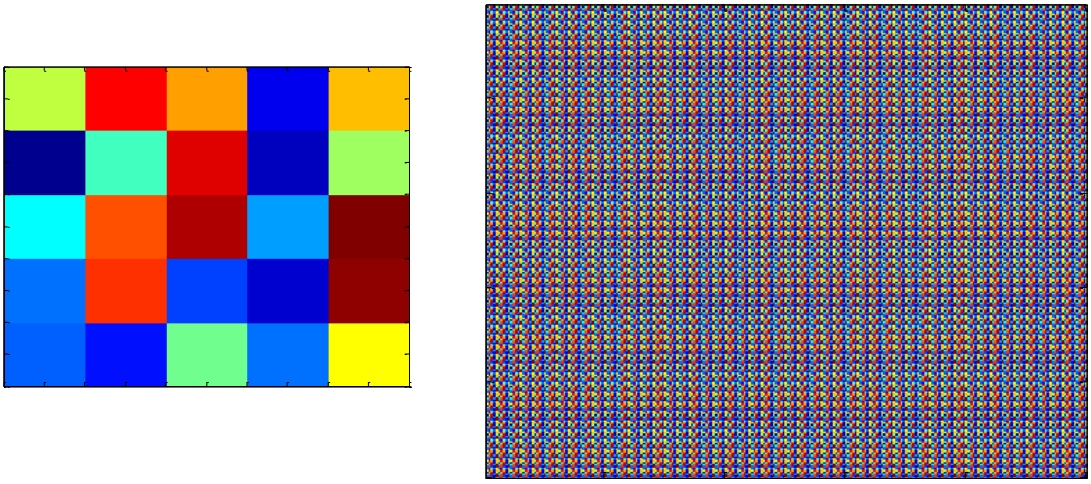


Figure 5: Examples of spectral Frames (one band per frame)

While the Hyperion utilizes a traditional architecture in order to acquire the spectral data, the IMEC mosaic relies on a specialized spectral filter pattern. Table 33 (left)

shows a 5x5 spectral sampling pattern utilized for acquiring a 25 spectral band imagery, while the right image presents the full resolution (200x200 pixels) of the spectral pattern, generated by repeating the 5x5 pattern along both horizontal and vertical dimensions.

Table 3: Spectral Filter pattern



The simulated frame that would be acquired by the IMEC mosaic camera is shown in Figure 6.

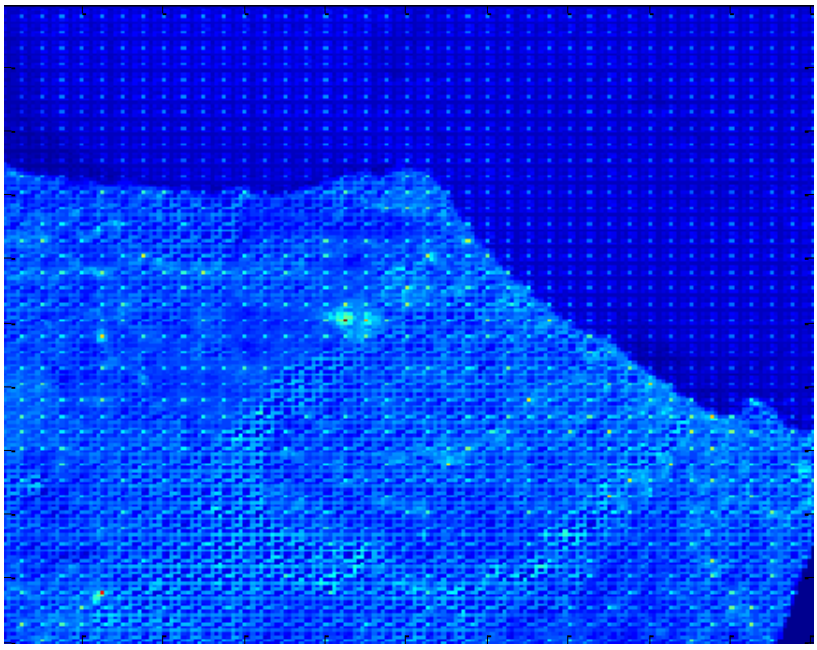


Figure 6: Simulated spectral mosaic frame

3.2.3. Data Sharing

Data sharing will take place using the PHySIS Filestote and will be restricted to the PHySIS consortium members only through a password policy. The reason for the access

limitation is that the primary purpose of this dataset is for internal use in simulation experiments.

3.2.4. Archiving

No long term archiving is planned for this particular dataset.

3.3. Laboratory Snapshot Mosaic Data

3.3.1. Description

This dataset will consist of data collected using the IMEC spectral cameras under laboratory conditions. An initial version of this dataset consists of i) multispectral images in the spectral range 600-875nm captured by the IMEC Mosaic camera and ii) hyperspectral data acquired by the Linescan system in the range 400-1000nm.

3.3.2. Standards & Metadata

The dataset contains imagery from an object (steak) acquired under controlled laboratory conditions. The dataset is composed of:

- Meat-Mosaic_600-875nm_Cube_Reflectance Correction: HYP cubes in the range 600-875nm obtained using the IMEC mosaic sensor, with reflectance correction applied.
- Meat-Mosaic_600-875nm_Cube_Spectral Correction: HYP cubes in the range 600-875nm obtained using the IMEC mosaic sensor. Both reflectance and spectral correction has been applied to this data.
- Meat-Mosaic_600-875nm_Frame-PGM-8bit_Irradiance: raw 8bit frames obtained from the IMEC mosaic sensor, without reflectance correction.
- Meat-Mosaic_600-875nm_Frame-PGM-8bit_Reflectance Correction: raw 8bit frames obtained from the IMEC mosaic sensor with reflectance correction applied.
- Meat-Reference_Linescan_400-1000nm_Cube_Reflectance Correction: reference HYP cube obtained using SpecIM camera. The meat scanned was different from the one used for scanning with our mosaic sensors. We have verified that the reflectance spectra is similar across the two meat samples.
- Meat-Reference_Spectrometer: the folder contains the reference spectra obtained using a spectrometer
- Mosaic_600-875nm_sensor_calibration_files: the folder contains following files for the mosaic sensor used, calibration file in XML format, filter responses in CSV format (This file contains the filter response for each band in the sensor for the 400-1000nm spectral range), the ‘central_wavelength.xlsx’ file containing the central wavelengths for different bands that was used for the cube construction
- RGB Image folder containing RGB image of the object which is a piece of raw meat

3.3.3. Data Sharing

Data sharing will take place using the PHySIS filestore to the members of the consortium. The PHySIS filestore employs a password protection in order to limit access to unauthorized personnel. The data are encoded in Matlab based files (*.mat files) that can be loaded using the associated Matlab script that will be included.

Public dissemination will be achieved through the PHySIS public website, as well as the institutional websites from consortium members including SPL-FORTH. In this scenario, only a limited part of the dataset will be freely available, under explicitly stated legal conditions concerting their use and the necessary citations.

3.3.4. Archiving

Long term archiving will be sought using publicly available platforms, including the SPL-FORTH website, the UCI machine learning repository and the OpenAIRE platform. The utilization of these platforms will support the long lasting dissemination of the dataset, free of charge. In general, the limited datasets that will be considered for long-term storage will be in the order of a few Gigabytes or less, thus removing the need for highly specialized solutions.