THE EARTH OBSERVATION IMAGE LIBRARIAN (EOLIB): THE DATA MINING COMPONENT OF THE TERRASAR-X PAYLOAD GROUND SEGMENT

D. Espinoza-Molina[1]*, V. Manilici*[2]*, O. Dumitru*[1]*, C. Reck*[2]*,*
S. Cui[1]*, H. Rotzoll*[2]*, M. Hofmann*[2]*,G. Schwarz*[1]*, M. Datcu*[1]
[1]German Aerospace Center(DLR), Remote Sensing Technology Institute (IMF), Munich, Germany
[2]German Aerospace Center (DLR), German Remote Sensing Data Center (DFD),Munich, Germany

ABSTRACT

In this paper we present the Earth Observation Image Librarian (called EOLib) as a new generation of Image Information Mining Systems. EOLib is operated in the Payload Ground Segment of TerraSAR-X. The main goal of EOLib is to provide semantic annotations of satellite image content and offer to the end user a semantic catalogue via a web user interface. Moreover, EOLib has more functionality such as searches based on image metadata and semantics, visual exploration of the image archives, metadata extraction, etc. The system consists of components such as a query engine, knowledge discovery in databases, visual data mining, epitome generation, and user services. EOLib is able to ingest a TerraSAR-X scene with 8000$\times$8000 pixels in about three minutes. The EOLib workflow starts with the ingestion of a scene, it continues with the semantic annotation of the image content based on machine learning methods, and it ends with publishing the semantic catalogue and enabling the search by metadata and semantic image descriptions.

***Index Terms*—** Software Architecture, Image Information Mining, Payload Ground Segment, Data Mining, Earth Observation.

1. INTRODUCTION

Over more than 15 years, major efforts have been made to introduce Image Information Mining (IIM) to Earth Observation (EO) data holdings in order to allow various applications to retrieve all desired information. These efforts have resulted in a major scientific progress in the understanding remote sensing specific principles of IIM, by developing and validating algorithms, methods and systems for EO knowledge discovery and data mining, by evaluating appropriate methods and technologies [1], by analysing the user needs and expectations, and finally, by successfully demonstrating the capabilities of IIM on EO data. These efforts resulted in several IIM system implementations as - for example - the Knowledge-driven content-based Information Mining system (KIM) [2] that presented the concept of image retrieval based on content using a Bayesian approach for the computation of similarities between images. Later, GeoIRIS (Geospatial Information Retrieval and Indexing System-Content Mining) [3] proposed a primitive feature extraction based on patches instead of pixels and semantic modelling. However, later, the problems of matching the image content with semantic definitions adopted by humans became more and more evident, causing the so-called semantic gap [4]. The semantic gap demonstrated the necessity of semantic definitions to be included in the image retrieval. In an attempt to reduce the semantic gap, more systems including labelling or the definition of the image content by semantic names were introduced. For instance, [5] demonstrated that the semantic representation has an intrinsic benefit for image retrieval by introducing the concept of query by semantic example (semantics and content). Later, in addition to the semantic definitions, the combination of several types of data in order to improve the image retrieval was presented in [6].

The European Space Agency (ESA) has been contributing to the development of IIM systems by funding several projects in this field. Among various system implementations, we have to mention the Knowledge-centred Earth Observation (KEO) system prototype. This system permitted users to interactively extract relevant features and information from EO data, and to provide outputs as, for example, valuable information extracted from data, in easily accessible formats. A second example is the KLAUS project, the main goal of which was an overall improvement of the KEO prototype system with a focus on having models for land use management. Currently, in this context and contributing to the big data era, where new mining techniques are necessary due to the volume, variability, and velocity of such data [7], we present EOLib (Earth Observation Image Librarian) as a next-generation of an Image Information Mining system implementing novel techniques for image content exploration and knowledge discovery in databases. EOLib produces information about the content of EO products which is usually hidden in raster data and metadata. EOLib is being operated in the Payload Ground Segment (PGS) of TerraSAR-X (TSX) and will aim at enlarging the IIM scope for a more complete exploration of the EO data sources by establishing large scale information mining functions within the multi-mission PGS operating missions like TerraSAR-X/TanDEM-X, Sentinel-1/2 and similar high resolution Synthetic Aperture Radar (SAR) and optical imaging missions. The EOLib system will allow users to find EO products of interest for their specific applications with semantic concepts. The EOLib system offers functions such as data model generation by means of tiling and the extraction of primitive feature from EO products, visual data mining to browse the image archives, knowledge discovery in databases to define a semantic annotation of the image content, queries based on different parameters via user services, and an epitome production functionality. Thus, EOLib will consist of a set of tools for sustainable long term and efficient utilization of EO data content. Our previous publication [8] discussed the planned EOLib architecture prior to implementation. We had the opportunity to review the architecture in ways that were not foreseeable before implementation. Interfaces were refined (both the connection graph and the interface semantics). Finally, we addressed several critical performance problems.

The rest of the paper is organized as follows. Section 2 presents an overview of the EOLib system architecture. Section 3 describes some examples and Section 4 concludes the paper.

doi: 10.2788/854791

2. EOLIB ARCHITECTURE OVERVIEW

The EOLib baseline architecture is depicted in Fig. 1. EOLib consists of several independent systems that communicate with each other via established interfaces in a service-oriented architecture. Each of these systems may integrate one or several components which provide specific functionality to the system. The novelty in EOLib is that there are several user-oriented components that enlarge the functionality of the Payload Ground Segment (PGS) such as Visual Data Mining (VDM), Knowledge Discovery in Databases (KDD), Query Engine (QE) and the User Services (US). The current version of the EOLib implementation operates with TerraSAR-X L1b products. The rest of this section briefly describes the PGS components represented in blue and the new EOLib components marked in orange.

2.1. Payload Ground Segment Components

The PGS of the German Aerospace Center (DLR) consists of several subsystems such as data acquisition, data processing, data and information management, librarian and archiving. EOLib is designed to be integrated mainly with the Data and Information Management System (DIMS) [9].

In the baseline architecture presented in Fig. 1, the PGS components are shown in blue and the data items in green. All DIMS components can be managed through a single interface, the *Operating Tool*. The *Processing* component coordinates the repeated operation of the underlying EOLib data model generation on batches of EO data. The *Long-Term Archive* (LTA) provides persistent storage and access to EOLib data items. The LTA is based on the DIMS Product Library (PL) and is the master repository for EOLib. An additional high-speed database is required for performance reasons as it shall be possible to rebuild the database from LTA data. Data from the processing workflow is uploaded via the *Ingestion Interface* to the *Data Mining Database* (DMDB) and EOLib data items are transferred and stored in the Long-Term Archive together with the standard L1b products. A simple mechanism that ensures data consistency (LTA against DMDB) is provided. The *Online User Services*, based on the PGS librarians' Earth Observation on the Web (EOWEB), EOWEB GeoPortal (EGP) and Geospatial Data Access System (GDAS), is enhanced by additional EOLib components for browsing a specific part of the DMDB like the semantic catalogue and by offering queries based on metadata and semantic annotations.

2.2. EOLib System Components

The new EOLib system components contain most of the novel and innovative EOLib functionality. They are shown in the baseline architecture in orange and the data items in yellow. Here it can be seen that DMG and DMDB are internal components managed by a tailored operator while the QE, KDD, and VDM components provide front-end functionality to the user and operator. The EOLib system modules are either independent components or functionality integrated into existing PGS components.

Fig. 1 shows that the *Data Model Generation* (DMG) is controlled by the *Processing* component. The data model generation component is a processing chain that produces EOLib data items from a TSX product and its metadata. The main DMG functionality is metadata extraction, image tiling with multiple resolutions, basic feature extraction, and high resolution quick-look generation. The output of DMG consists of metadata, tiles, high resolution quick-looks, and extracted features. They are saved into the LTA. These data are later transferred to the data mining database via the *Ingestion Interface*. Currently, DMG counts on three feature extraction methods, namely, Gabor Linear Moments (GLM) [10], Gabor Logarithmic Cumulants (GLC) [11], and Weber Local Descriptors (WLD) [12]. These methods are tile-based and extract mainly texture as image descriptors. The *Data Mining Database* (DMDB) provides high-speed storage and some data mining functionality whose processing and retrieval performance requires a database-close implementation. It is based on the relational database MonetDB [13]. The DMDB component manages data handling, storage, administration and some of the processing for the entire EOLib components.

The *Query Engine* (QE) allows the user to search for EOLib relevant data. The following types of queries are supported: 1) Querying based on metadata, where the user can query the image archive using standard metadata as, for example, coordinate systems, type or product, acquisition time, etc. 2) Querying based on semantic annotations; here the user can select a semantic label from the available labels in the semantic catalogue to perform the query. It is worth to mention that those labels are pre-defined labels previously obtained as results of the semantic annotation by using the Knowledge Discovery tool. The *Knowledge Discovery in Databases* (KDD) component adds semantic annotations to EO products. It consists of a Graphical User Interface (GUI) which interacts with the user and receives the user's input; it is based on relevance feedback methods and the KDD core which accepts the user's requirements from the GUI and uses them to get the data from the DMDB. The KDD core component includes a Support Vector Machine (SVM) as its machine learning method in order to classify the image content and to define semantic labels. In the semantic definition step, the operator's feedback is passed as training data to the machine learning method and it performs the prediction of the results [14]. After the training, the whole data can be annotated. The *Visual Data Mining* (VDM) component is a structured browsing facility for large amounts of image data. It is composed of a GUI which allows the user to navigate in the image archive. It gets the data from the DMDB and adapts them by dimensionality reduction to present them to the user through the GUI [15]. The *Epitome Generation* (EG) is integrated with the DMDB. The epitome is a summary of the EO product information content (i.e. metadata, high resolution quick-looks, basic features and all the annotations) as presented as actionable information for information mining in individual EO products. The epitome is a result of the data model generation and semantic annotation and may be delivered with the standard EO product or as a distinct product component. It is intended to be used for the individual product content inspection. Epitome access is done offline, on the user's PC/notebook, using the Epitome Browser (a separate program).

3. EXPERIMENTAL RESULTS

The DMG of the EOLib system starts the ingestion of the TerraSAR-X product. It sets the input parameters (i.e., product path, patch size, levels of resolution, etc.). Later, during the metadata extraction, the XML annotation file is read and it extracts the relevant metadata entries as, for example, the four corner coordinates, acquisition angles, resolution, pixel spacing, number of bands, acquisition time, etc. Further, the TSX image is tiled into several patches generating a grid of multi-size patches together with their high resolution quick-looks. Later, the primitive features are extracted from each generated patch by the selected methods. Finally, all the generated information is written into an XML file called the data model and it is transferred to the long term archive. In a further step, the ingestion interface uploads the generated information into the DMDB,

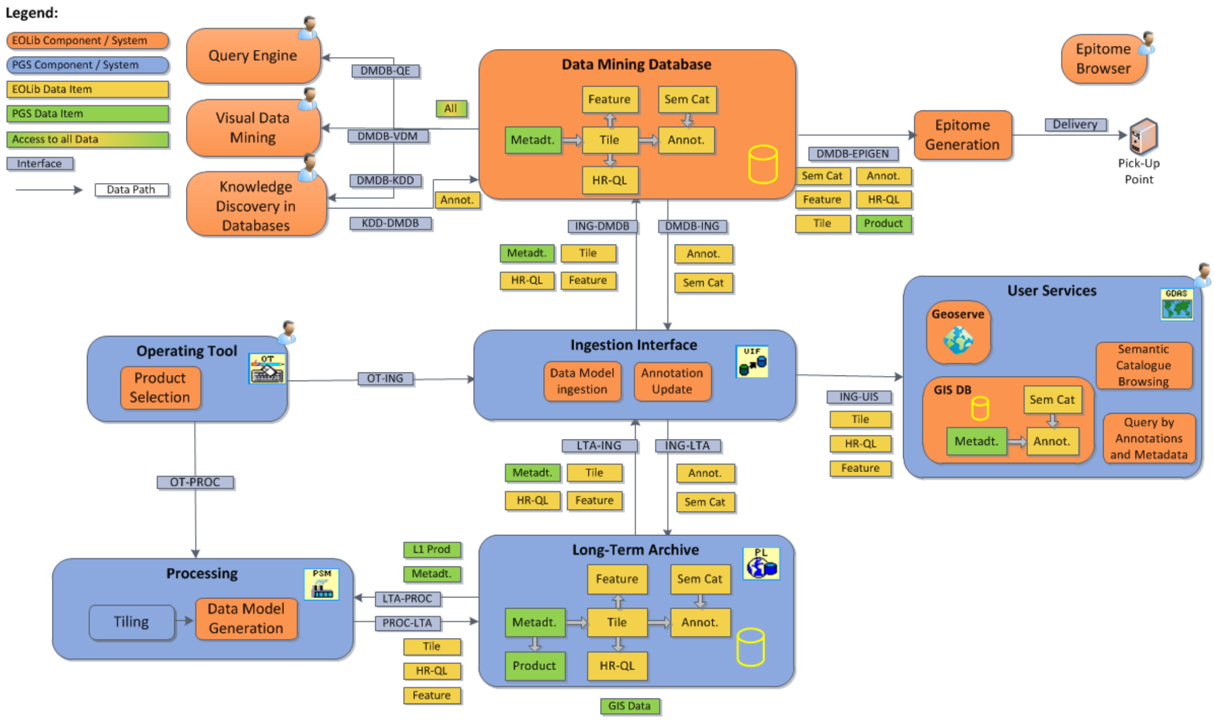

Fig. 1: EOLib Baseline Architecture. Systems/components are shown as rectangles with rounded corners (blue for the current PGS and orange for new EOLib components). Interfaces are displayed as arrows pointing to the direction of the information flow. Components with user interfaces have an icon user shown in the upper right corner. Information items transmitted via an interface are displayed near the interface arrow as rectangles (yellow for EOLib and green for PGS).

thus enabling the remainder of the components (i.e., KDD, VDM, QE). The generation of the data model using a TerraSAR-X scene of 8000×8000 pixels takes approximately less than three minutes, which is a reasonable computing time in the big data era. The use of metadata enriches the data model by adding more parameters that can be later used in advanced queries. The data model will be completed by adding semantic annotations of the image content provided by active learning methods. The next step is to provide semantic definitions to the image content. This function is performed by using the KDD component, which allows us to load one or several scenes and to annotate the image content. These annotations are stored back into the DMDB and will be transferred later to the user services via the ingestion interface. The US will allow queries based on metadata and semantic descriptions via a graphical web user interface.

The EOLib system has been processing mainly TerraSAR-X scenes. Currently, the DMDB comprises about 1200 scenes taken from around the world. These scenes were tiled on two grid levels with sizes of 256×256 and 128×128 pixels resulting in approximately ten million tiles. From each tile its primitive features were extracted using Gabor filters (GLM and GLC) with 4 scales and 6 orientations as input parameters, and a Weber local descriptor with 18 excitation levels, and 8 orientations as input parameters. Both Gabor methods yield a primitive feature vector of 48 dimensions, and a vector with 144 dimensions in the case of Weber descriptors. As examples of semantic definitions we annotated several scenes with land use and land cover categories. The categories were taken from the EO Taxonomy presented in [16], which is hierarchically organized in two levels. The first level includes 8 main land cover land use classes like like urban areas, transport, industrial areas, agriculture, military facilities, bare ground, water bodies, and natural vegetation, while the second level contains about 10 subcategories for each main category. In urban area, for example, we can find semantic categories like high buildings, high density residential areas, informal settlements, low density residential areas, skyscrapers, etc. Fig. 2 shows an example of the GUI used for semantic definition and Fig. 3 displays the user web services interface.

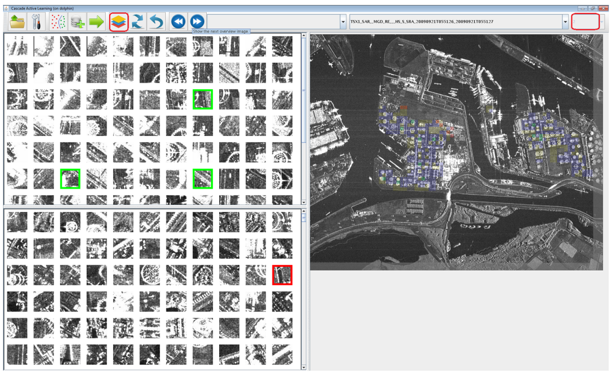

Fig. 2: Example of the Knowledge Discovery GUI component. The tiles marked in blue represent oil containers that were found using machine leaning methods implemented in this component.

In the example shown in Fig. 2, a TSX scene is displayed on the right part and a list of its tiles is shown on the left part. We can observe that the scene has categories like oil containers, water bodies,

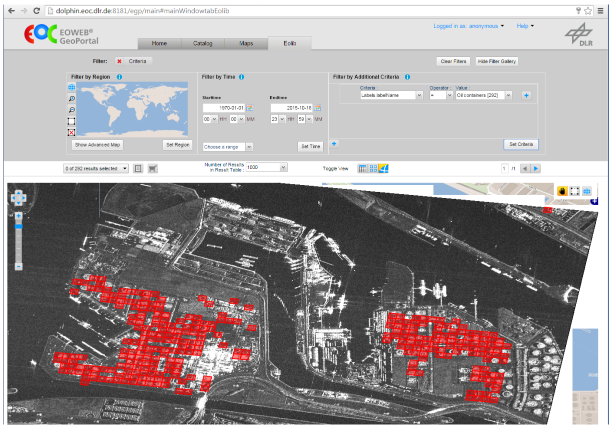

Fig. 3: Example of the User Web Services component. The tiles marked in red represent oil containers that were found using machine learning methods.

industrial areas, etc. In this example, the operator is looking for oil containers, so the tiles marked in green are the positive examples, while the negative examples are marked in red. Those examples are passed to the SVM. The SVM performs the prediction of the desired label and returns the classification result, which is presented in blue. When the classification is satisfactory, the annotation of the tiles with semantic categories is stored in the DMDB and further transmitted to the LTA and then this information will be available for searches in the US. The US interface is shown in Fig. 3; here it can be seen that a search for "oil containers" has been made. The results are highlighted in red.

4. CONCLUSIONS

In this paper we introduced the Earth Observation Image Librarian (EOLib), the data mining component of TerraSAR-X being installed in the TerraSAR-X Payload Ground Segment. The implemented system is installed in the TerraSAR-X Payload Ground Segment. The architecture is presented as a modular system integrating several components with well-defined functionality allowing the improvement or replacement of each component without affecting the whole system. It provides user services such as searches based on different parameters as, for example, metadata and semantic annotations of the image content. The ingestion of a new scene takes on average three minutes.

5. ACKNOWLEDGMENT

We would like to thank ESA for funding the EOLib project.

6. REFERENCES

[1] O. Dumitru and M. Datcu, "Information content of very high resolution SAR images: Study of feature extraction and imaging parameters," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 8, pp. 4591–4610, Aug. 2013.

[2] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P.G. Marchetti, and S. D'Elia, "Information mining in remote sensing image archives: system concepts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 12, pp. 2923 – 2936, Dec. 2003.

[3] C.-R. Shyu, M. Klaric, G.J. Scott, A.S Barb, C.H. Davis, and K. Palaniappan, "GeoIRIS: Geospatial Information Retrieval and Indexing System-Content Mining, Semantics Modeling, and Complex Queries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 839–852, 2007.

[4] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000.

[5] N. Rasiwasia, P.J. Moreno, and N. Vasconcelos, "Bridging the Gap: Query by Semantic Example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923 –938, Aug. 2007.

[6] D. Espinoza-Molina and M. Datcu, "Earth-Observation Image Retrieval Based on Content, Semantics, and Metadata," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 11, pp. 5145–5159, Nov. 2013.

[7] A. Bifet, "Mining big data in real time," *Informatica (Slovenia)*, vol. 37, no. 1, pp. 15–20, 2013.

[8] V. Manilici, S. Kiemle, C. Reck, and M. Winkler, "Architecture concept for an information mining system for earth observation data," in *Proc. PV*, 2013.

[9] M. Wolfmüller, D. Dietrich, E. Sireteanu, S. Kiemle, E. Mikusch, and M. Böttcher, "Data Flow and Workflow Organization - The Data Management for the TerraSAR-X Payload Ground Segment," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 44 –50, Jan. 2008.

[10] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 837–842, Aug. 1996.

[11] J. Singh and M. Datcu, "SAR Image Categorization With Log Cumulants of the Fractional Fourier Transform Coefficients," *IEEE Transaction on Geoscience and Remote Sensing*, vol. 51, no. 12, pp. 5273–5282, 2013.

[12] S. Cui, C.O. Dumitru, and M. Datcu, "Ratio-Detector-Based Feature Extraction for Very High Resolution SAR Image Patch Indexing," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 5, pp. 1175–1179, Sep. 2013.

[13] CWI, *Monetdb Copyright (c)*, Mar 2008-2011, Available from: http://www.monetdb.org/Home.

[14] P. Blanchart, M. Ferecatu, S. Cui, and M. Datcu, "Pattern retrieval in large image databases using multiscale coarse-to-fine cascaded active learning," *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, vol. 7, no. 4, pp. 1127–1141, April 2014.

[15] D. Espinoza-Molina, M. Datcu, D. Teleaga, and C. Balint, "Application of visual data mining for earth observation use cases," in *Proc. ESA-EUSC-JRC 2014 - 9th Conference on Image Information Mining Conference: The Sentinels Era*, 2014, pp. 111–114.

[16] C. O. Dumitru, S. Cui, G. Schwarz, and M Datcu, "A Taxonomy for High Resolution SAR Images," in *Proc. ESA-EUSC-JRC 2014 - 9th Conference on Image Information Mining Conference: The Sentinels Era*, 2014, pp. 89–92.