Architecture Concept for an Information Mining System for Earth Observation Data

Vlad Manilici, Stephan Kiemle, Christoph Reck, Mario Winkler

German Aerospace Centre (DLR)
Münchner Straße 20, 82234 Weßling, Germany
Email: {Vlad.Manilici, Stephan.Kiemle, Christoph.Reck, Mario.Winkler}@dlr.de

ABSTRACT

EOLib, the *Earth Observation Image Librarian* is an upcoming *Image Information Mining* (IIM) system for *earth observation* (EO) products. As integral part of a *payload ground segment* (PGS) it operates on the original EO products, metadata, and on computed higher level abstractions including basic features and semantic annotations of product tiles. The core goal of EOLib is to introduce mature information mining functions in existing EO payload ground segments.

EOLib is integrated with the multi-mission PGS existing at DLR's premises. Operations including product tiling, feature extraction and automatic annotation are performed within the PGS services infrastructure. Intermediate data including features and quick look images are forwarded to the EOLib core to perform additional data mining operations as: semantic learning, content-based information retrieval and visual data mining. The PGS user services are augmented with query engines for semantic annotations and metadata and with a semantic catalogue browser.

We present in this paper the architecture concept of the EOLib system. It starts with a set of functional requirements, the constraints imposed by the existing PGS and the experience gathered from prototype standalone IIM systems. System components are subsequently identified based on logical functionality blocks. Data flows and interfaces are built in order to allow simple, clear system integration and to maximize performance. We conclude with the implementation of a few use cases.

Keywords: Software Architecture, Image Information Mining, Integration, Payload Ground Segment

INTRODUCTION

The *Earth Observation Centre* (EOC) in the *German Aerospace Centre* (DLR) provides *earth observation* (EO) research and development activities, as well as operational tasks for data reception, processing and archiving. EOC runs its own multi-mission *payload ground segment* (PGS) for receiving, processing and the storage of petabytes of EO satellite data, including high resolution radar and optical data.

EO data users are challenged through the continuous increase in data resolution and size, as shown in Figure 1. The effect is that EO archives are growing asymptotically in size and information content. Finding relevant data is restricted by the limitations on the searching capabilities of present archives, being purely based on properties such as sensing time, geographical location, processing level and other descriptive product metadata. This raises the necessity for new and efficient search tools relying on image structure and content.

Image Information Mining

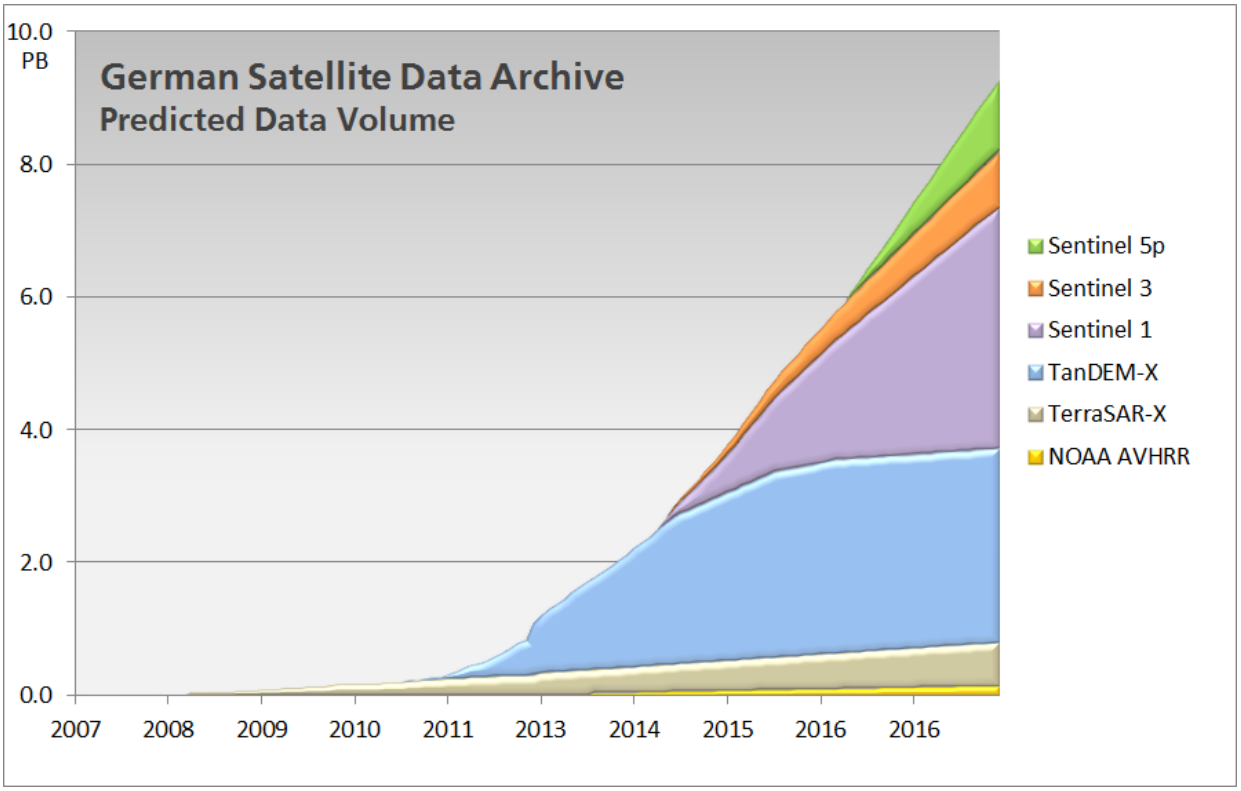

Figure 1 Predicted volume of the German Satellite Data Archive operated at EOC

According to [1]: "*Content-based image retrieval* (CBIR) systems allow searching for images in a large database when the image content is queryable. Initially the term content referred to colors, shapes, textures, or any other information that can be derived from the image itself. Presently «content» refers to the semantic description of the imaged structures, and specifically in the earth observation domain, to some of their physical parameters. Generally, CBIR systems are composed of several modules such as feature extraction, clustering, catalogue generation, semantic definition, and relevance feedback."

A new field of study arose lately: *Image Information Mining* (IIM), which integrates concepts of CBIR, *Knowledge Database Discovery* (KDD) and data mining. "An IIM system allows the user to access a large image repository, to extract and to infer knowledge about the patterns hidden in the images by retrieving dynamically a collection of relevant images and provide high levels of abstraction by defining semantics. Thus it enables the communication between heterogeneous sources of information and the user with diverse interests at semantic" level. This excellent definition was given in [1].

The EOLib Project

The desire to improve the usability of the EO data archives led during recent years to increasing efforts towards the introduction of IIM technologies [2]. These studies improved the understanding of the appropriate methods, algorithms and technologies together with the user needs and expectations.

In order to perpetuate this innovation process, ESA and DLR are conducting the *Earth Observation Image Librarian* (EOLib) project in order to:

- Analyze IIM concepts and methods for mining heterogeneous sources as EO images, their metadata, Image Time Series, and other related geo-information;
- Select, develop and elaborate state-of-the-art algorithms and tools up to operational maturity;
- Provide a modular architecture with clear interfaces allowing the incremental implementation and integration of processing and data mining components;

- Integrate high-performance IIM into existing multi-mission payload ground segments and user services by extending the existing long-time archive and catalogue to adequately access and preserve high-resolution quick looks, features and semantic annotations of tiles;
- Demonstrate IIM with large volume data sets, as TerraSAR-X, the Sentinels and similar high resolution SAR and optical missions. EOLib shall provide the processing performance, I/O and storage capacity for:
    - tiling and feature extraction of basic image content;
    - generation of high-resolution browse images required for visual inspection;
    - generation of semantic annotations;
- Identify unusual and yet undiscovered patterns in large EO datasets and time series;
- Provide in the long-term a tool for sustainable long term and efficient content-discovery and utilization of very large archives of high-resolution EO data.

EOLib is currently undergoing the detailed design and a prototype implementation and integration.

EARTH OBSERVATION IMAGE LIBRARIAN

Usage Scenarios

A fundamental part of the EOLib concept is a semantic search engine, which is integrated as an organic component within an EO PGS system containing a large EO data archive. Since EOLib aims at interpreting image content and associating it with other information in the background, the search engine is the visible part of the system interacting with the user e.g. to refine his expression of needs, to inform him about data content and to make suggestions about appropriate products including alternative interpretations. In this section we present an EO data retrieval scenario using the EOLib search and query engine which operates on semantic annotations and descriptors trained within other parts of the EOLib system.

The EOLib search engine provides content-based filter possibilities which can be combined with classic product metadata criteria. The user can enter a semantic label in the form of text or select an item from the available labels in the catalogue to perform the query (e.g. skyscraper, storage tank, bridge or hotel resort). These labels are previously obtained as results of the image annotation process. Also the user can compose more complex query expressions using these labels along with semantics and topological relations (“north of”, “contained in”) and numerical operators (e.g. “less than”). The interface also allows to extract quantitative statistical information e.g. extent of industrial areas, commercial, residential, vegetation, etc. in percentages or in surface units for a specified inhabited area.

The retrieved data can be used in a decision making process, facilitating its integration in other types of models, for e.g. an estimate of the quality of life. Also, image time series over a specific geographical area, can be used for environmental monitoring e.g. to observe the growth of mega cities or fast growing urban areas in developing countries to monitor the evolution of urbanization. Also EOLib can support *ad hoc* rapid mapping activities to help faster and more complete generation of risk, damage, or situational assessment maps to provide a systematic monitoring of large geographical area affected by a disaster. The system contributes to this scenario by means of content-based fast image or image time series retrieval supporting the relevant data selection and delimitation within regions of interest.

Data Flow

As the EO data is ingested into PGS and stored in the *Long-Term Archive* (LTA) the EOLib system analyzes the received images and their associated metadata. At first the EO image is divided into tiles of different scales and resolutions and high-resolution quick-looks are created for each tile. A tile is a part of the image, which can contain one or several objects depending on its size.

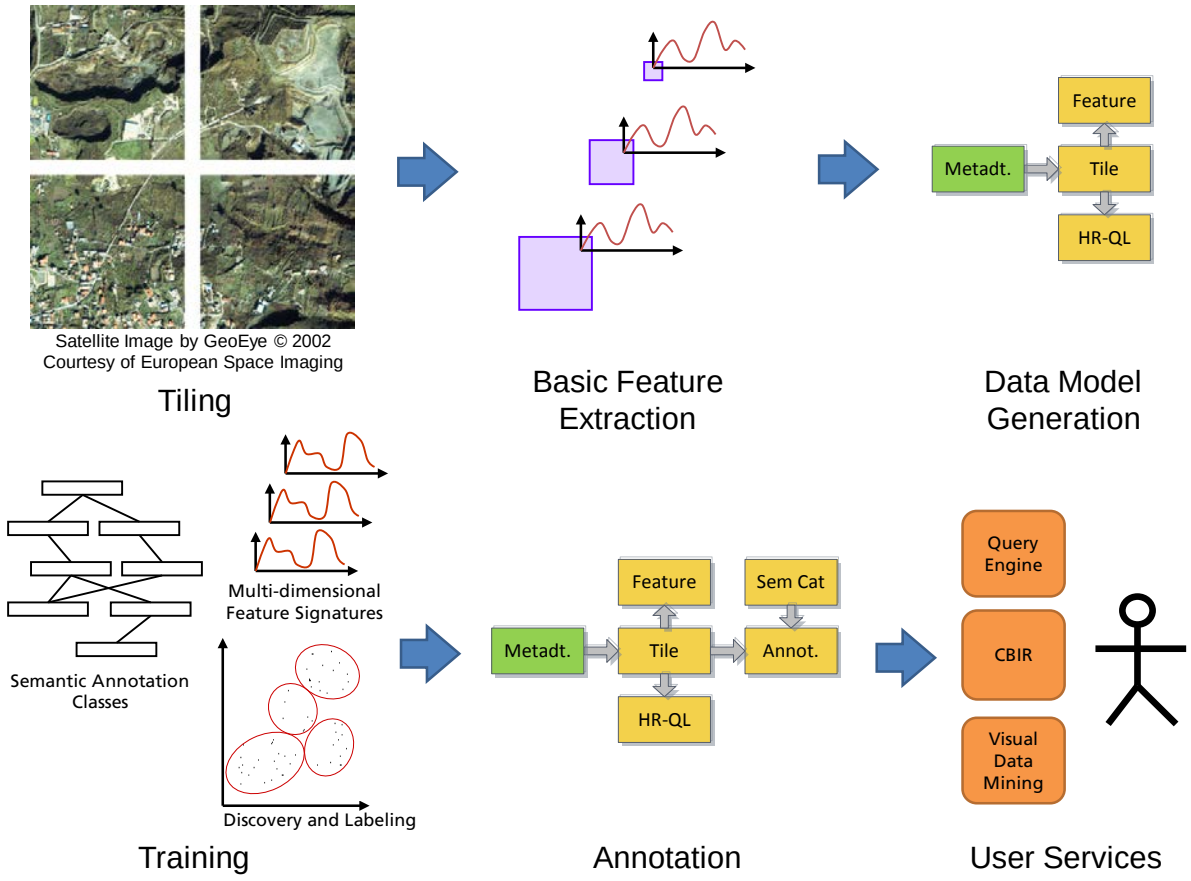

Figure 2 Dataflow of for Image Information Mining

In the next step a set of basic features is computed from each tile to generate descriptors of its content. Basic features extract characteristic information from the tile data by applying a mathematical function on and reducing the amount of tile data. Examples are: Gabor filters [3] and Weber Local Descriptors [4].

Once this is completed, data mining functions will be applied to the sets of basic features to discover semantic annotations describing the content of the tile e.g. skyscraper, storage tank, bridge or hotel resort. The annotations are stored into a semantic catalogue, which represents a taxonomy of semantic labels together with the ontological relations between them. The next step is the generalization of the induced semantics to all image products by a continuous background annotation process.

EOLib applies clustering algorithms in a multidimensional space to perform the semantic annotation of the data, supporting the use and generation of expert knowledge and using elements of an extensible taxonomy. The definition of the spatial-temporal context can be further extended by adding auxiliary GIS data (e.g., CLC 2000 [5], Urban Atlas [6], DEMs, or street maps) to EOLib and applying the data mining techniques to them.

Finally a data model is created which will be an *Epitome* of the image content, i.e. a compact representation of classes of scene structures or their evolution. The *Epitome* is the basic component for:

- An enriched index for the product catalogue,
- A new type of interactive value added EO products,
- The image and data representation for further auto-annotation and data mining processing.

Further, the EOLib system will help the human operator, in a separate interaction loop, to use mining functions for the aggregation of the obtained image and data elements and also to add semantics with alternatives depending on the application context.

Performance Budget

Performance is paramount in order to achieve a usable IIM system that operates on a relevant amount of EO data in a timely manner. The system shall be able to process the input data in a reasonable amount of time, in the magnitude of several weeks. Feature discovery and queries shall be processed interactively, in a matter of seconds or at most minutes. Moreover, EOLib usage is not allowed to disrupt the ongoing PGS operation.

At this early project phase, in order to design the architecture, we can only estimate the required performance based on the anticipated amount of data and required bandwidth. Data that is accessed often is stored locally and unnecessary data transfers are avoided. Dedicated interfaces and proven technologies are used whenever possible.

The first EOLib evaluation uses the data set from the TerraSAR-X [7] radar satellite mission, specifically of the type: high resolution single polarization spotlight, radiometrically enhanced, *Multi-look Ground-range Detected* (MGD) products [8]. A typical product with the image data, metadata and quick looks has a size of about *300 M*B and covers about *25* square km. To estimate the required performance, we assume an example region of the size of Bavaria with about *850* GB of data, and the need to ingest the data within one week, which leads to an ingestion processing bandwidth of *5* GB/hour. The PGS can accept processing results (tiles, features, quick looks) with less than about twice the original amount of storage. These sizes and throughput are reasonably manageable by current technologies.

When automatically annotating the tiles, only the feature vectors need to be processed. Their size is only a fraction of the whole product. This allows for processing the example area with the same processing bandwidth as above (*5* GB/hour) within a few hours.

ARCHITECTURE

Concept

EOLib is built by integrating several image mining modules with the existing PGS [9] system implemented by the Data and Information Management System (DIMS) [10] and operated by ESA in its Multi-Mission Facility Infrastructure and by DLR in its multi-mission PGS. DIMS is inspired by the Service Oriented Architecture [11] paradigm. Its architecture is scalable, offers consistent interfaces and thus limits the amount of work required to extend the system.

A PGS based on DIMS consists of several services registered with a central name service. Services communicate using a standard interface protocol. Each service serves a well-defined function. Services are instantiated by configuring generic software to perform specific tasks. For example, we use a versatile processing management component for most data processing tasks. The DIMS software is highly configurable: it operates as a multi-mission EO facility (serving multiple satellite missions) and is deployed in several setups at different sites.

The goal of the EOLib project is to integrate the novel IIM components into an operational multi-mission PGS installation, without interfering with the on-going operations. Some of the IIM software was already prototyped as part of scientific projects at DLR; other components are new. We isolate IIM components according to atomic, well-defined functions. Some of the existing components are extended to use PGS standard service interfaces; other continue to use their (non-standard) interfaces; yet other are wrapped into a PGS management service. We augment the PGS with additional data mining functionality: ingestion processing, query by semantic annotations, *Epitome* generation etc. Data required for these functions, even if created by IIM services, is stored in the PGS *Long-Term Archive*, allowing it to be operationally used by the PGS. Some data mining operations, e.g. clustering and visualization remain in the realm of an *IIM core* consisting of the *Data Mining Database* and legacy software modules directly working on this data holding.

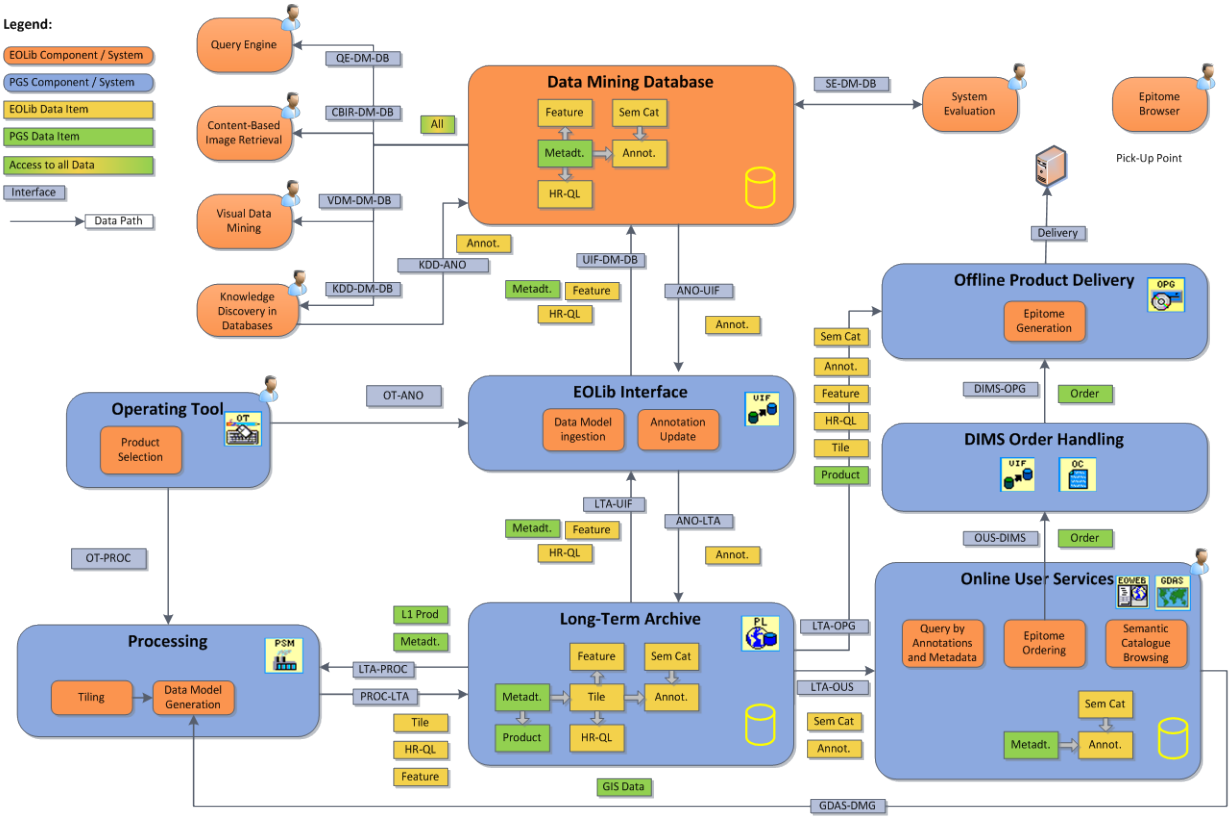

Figure 3 EOLib Architecture

Figure 3 shows the EOLib services and components, interfaces and data flows together with the most important data structures. IIM services are colored orange, PGS services blue, interface names grey, existing data green and novel data yellow. Arrows show interface connections and data flow directions. When appropriate, the icons of the DIMS software modules implementing the PGS service are shown in a corner.

A quick look reveals that EOLib consists of the *PGS Infrastructure* block and a separate *IIM core*, connected by the *EOLib Interface*. PGS services are augmented with additional configuration and components for EOLib.

PGS Infrastructure

Only the relevant PGS services contributing to the EOLib architecture are shown in the figure above. A part is depicted collectively as the *DIMS Order Handling* service. There is no central service managing workflows. Data and requests are tracked and processed in a sequence of interlinked but yet autonomously running services [10] [12].

The *Operating Tool* [13] is a special service that provides an operator interface which may connect to all services in a PGS installation. The OT user can follow requests in different services. Another special service is the *Long-Term Archive* [14] that provides persistent storage and retrieval for several petabytes of EO data. EO data is stored in the LTA at all processing stages and not directly transferred between PGS services. Any PGS service that needs EO data retrieves it directly from the LTA and, if necessary, stores back the processing results. Data inquiries are initiated by users with help of the *Online User Services* interface. The service is augmented with user interfaces for *product query*, *browsing the semantic catalogue* and ordering an *Epitome*.

PGS has several specialized processing services; EOLib adds an additional one for ingestion processing. The *Offline Product Delivery* builds and delivers data to the customers. OPG is augmented with functionality to build the EOLib *Epitome*. The *EOLib Interface* is a novel service based on an existing PGS service framework that manages the synchronization of the *Long-Term Archive* and the *IIM Core* with the *Data Mining Database*.

IIM Core

The *IIM core* consists of a high-performance database, the *Data Mining Database*, paired with several data mining services. Some data mining functions are implemented for speed purposes directly in the database. The *Knowledge Discovery in Databases* component manages the annotation rule discovery (on a sample) and the automated annotation process (on the whole collection). At this design stage it is the only component whose output flows into the LTA.

The *Query Engine* provides innovative user services to search by metadata and by semantic annotations, using labels available in the semantic catalogue. Special relation operators on the metadata will allow locating changes in time-series image data. *Content-Based Image Retrieval* provides a query-by-example functionality [15]. *Visual Data Mining* is visualization software for large image collections. Stable functionality will be taken over and integrated in the query component of the *Online User Services*.

REALIZATION

Data Model

Both within the PGS and the IIM components, EOLib uses and extends the pre-existing product data model with additional information needed for data mining. An EO product consists of metadata, describing the image (e.g. coordinates, instrument configuration, processing level) and the image (data take).The product components are stored in the *Long Term Archive* and can be retrieved for further processing or delivery.

We augment the product model with further elements in order to store *tiles, features, quick looks* and *semantic annotations*. The available set of semantic annotation values is described in an additional *semantic catalogue*. A product is associated with a multitude of tiles, building a resolution pyramid. Tiles are each associated with a quick look, a feature vector and several annotations. The augmented EOLib product can be delivered in its entirety for further study as an *Epitome*.

Interfaces

Within the PGS realm, data are transferred through the standard PGS remote invocation interface. IIM components integrated in the PGS use the interface provided by the overlaying PGS service. The *IIM core* features a database-centric architecture where all communication is done using the database interface. The two realms are connected by a synchronization service, the *EOLib Interface*. The EOLib Interface ensures that product metadata and the data computed during ingestion (*features, quick looks*) are transferred from LTA to the *Data Mining Database* (DM-DB) and the annotation results (*semantic annotations, semantic catalogue*) are transferred from the DM-DB to the LTA.

The highest data rates are expected on the LTA ingestion and *Data Model Generation* interfaces. Other interfaces are expected to provide a capacity of an order of magnitude lower as required to transfer the original EO products.

Ingestion

Selection of EO products for which the feature extraction shall be initiated is performed via the *Operating Tool*, the UI and control service of the PGS. A bulk processing command with a list of catalogue EO products is transferred to the *Processing* [16] service, responsible for coordinating the individual steps (tiling, data model generation). The results (tiles, quick looks, features) are uploaded to the *Long Term Archive* and metadata and features are further transferred to the *Data Mining Database*. The high resolution quick looks are also transferred to the DM-DB.

The Processing component is designed to wrap the *Tiling* and *Data Model Generation* components to provide request management functionality with the standardized interfaces of the PGS [10].

The *Long Term Archive* [14] stores a comprehensive catalogue of EO products from multiple missions spanning decades. We augment the product model with items required by EOLib: tiles, features, quick looks, semantic annotations and a semantic catalogue. The archive is designed to protect data against catastrophic loss due to hardware failure, misconfiguration, software or human errors. For this reason, the *Long Term Archive* is the authoritative EOLib repository and can be used if necessary to restore other duplicated databases.

Yet, the *Long Term Archive* might not offer the performance required for directly integrated data mining. For this reason, a certain level of data duplication is done in order to allow exploration of future data mining capabilities on the specialized *Data Mining Database*. Data mining results, such as annotations, are stored back in the *Long Term Archive*.

Feature Discovery

A training sample of the original tiles and features are subject to feature discovery and initial semantic annotation, with interaction from the operator.

The functionality is entirely contained within the *IIM core* realm. The required data results from ingestion processing and is made available in the *Data Mining Database*. Feature discovery and initial annotation are supported by the *Knowledge Discovery in Databases* service. Data is accessed over a native database connection. The resulting feature patterns and annotation definitions are held in the *Data Mining Database*, usable for the subsequent Background Annotation.

Background Annotation

After relevant features are discovered and annotated on training samples, the *IIM core* can proceed to annotate all the tiles in the data collection. The *EOLib Interface* transfers the feature information per tile from the LTA and ingests it into the *Data Mining Database*. This induces the annotation process in the *IIM core* which matches the trained feature patterns to the features of the tiles.

Upon match with a certain level of similarity, the corresponding annotation is attached to the corresponding tile with a numeric certainty measure. Several annotations may be attached to the same tile. The tiles may overlap to a certain degree to allow identification of patterns at the border of tiles. Therefore several tiles of the same product may obtain the same annotations.

Finally the new annotations are transmitted back through the *EOLib Interface* to the LTA, from which they are ingested with standard PGS product data/metadata upload functions into the *Online User Services*.

Annotation and Metadata Query

The PGS end-user front end, EOWEB, has new services added for querying for EO products on semantic annotations and metadata, browsing the semantic catalogue and ordering epitomes. Both free-text and input field queries are offered. The necessary data consisting of the semantic catalogue, semantic annotations and high-resolution tile quick looks is transferred to the EOWEB database.

Data Mining Operations

Using the *Data Mining Database*, several legacy components work directly with the extracted features and product metadata in order to explore further mining capabilities. This permits validating the extracted features, finding new features for a better separation of similar patterns in the feature space, evaluating the annotation process, and refining semantic search functions before integrating them in the *Online User Services*. As these operations are decoupled from the PGS architectural integration, they are not further elaborated here.

Epitome Generation

Epitomes are intended for off-line use. Users request them through the PGS user interface, the *Online User Services*. The order is passed through the PGS order handling chain and dispatched by the PGS delivery service, the *Offline Product Delivery*. The required product including the components generated in the context of EOLib are retrieved from the *Long Term Archive*, packed and delivered. This functionality is implemented similar to the product ordering workflow, using existing PGS components and interfaces. A dedicated *Epitome* browser can be used at the user premises as a client to display and explore the content of the delivered products.

DISCUSSION

This paper presents the current status of the EOLib architecture. The project passed the preliminary design review. Some unknown parameters that may influence the architecture in detail remain, for instance, lack of performance might force us to setup local storage or parallel processing. Yet, we expect the general concept to be implemented as described.

One interesting question is why we chose this architecture instead of other several possible solutions. Within DLR we could choose between the *Data and Information Management System* supporting product ordering of existing and future products and reprocessing, and the *Geospatial Data Access System* supporting OGC-compliant data access services for geographic data. The EOLib scenarios require systematic data processing functionality. For this reason, we decided to integrate it in a processing environment as opposed to a data access or data visualization system. Most of the EOLib processing steps have results for which permanent storage is sensible, as opposed to ad-hoc processing for visualization. Data mining functions in EOLib are either relatively simple (e.g. textual annotation queries) or complex enough (e.g. visual data mining) to preclude integration in an existing GUI concept.

The EOLib architecture concept defines how the IIM components are integrated within the PGS infrastructure. Some components are encapsulated in PGS services while others are interfaced as part of the *IIM core*, minimizing the amount of changes and integration work. Further integration of IIM components into the PGS is obstructed by the close coupling of the current prototype components with the DM-DB database. Keeping a separate *IIM core* presents the advantage of having a single and simple interface to the PGS. Avoiding further modularization and the creation of additional interfaces is a compromise between work load and system coherence.

Another compromise was made for data duplication. Duplication may harm due to storage size or if the copies become inconsistent with each other. On the other hand, duplicated data may improve performance due to the use of a local storage, faster than a remote one.

Apart from the *Long-Term Archive*, some of the relevant data is stored in the *Data Mining Database* and the *Online User Services*. Our data duplication decisions are justified by the following considerations:

- The speed of a local storage is needed in the *Online User Services* (holds copies of the catalogue, annotations etc.) and in the *IIM Core*. These services provide an interactive user interface and depend on quick access to data.
- Specialized databases are sometimes hard to use for general-purpose operations. With EOLib data held both in the LTA and in the DM-DB, EOLib profits from their individual advantages:
  - LTA is specialized for secure, long-term storage of a large (multi-petabyte) archive. It offers best-effort performance through a caching mechanism that uses magnetic discs for quick access and data tapes for bulk storage. The LTA has been designed for high availability, to store data from multiple missions, and its data-model is self-describing and open for evolutions.
  - DM-DB is a disk database, residing on a single partition and specialized in delivering speed to a certain class of queries.

In the next project phase we will incrementally build the PGS-integrated EOLib demonstrator, going from prototype components, over an integrated prototype, towards an integrated system. EOLib will be subject to software quality and user acceptance assessments. This incremental approach allows adaptation of the proposed architecture in order to match accumulated knowledge. We are confident that the proposed solution satisfies the system requirements, is elegant and implementable in a reasonable amount of time.

ACKNOWLEDGEMENTS

We thank the EOLib project initiator Prof. Dr. Mihai Datcu and ESA for managing the contract within the Earth Observation Program of the ESA Directorate.

REFERENCES

[1] D. M. Espinoza, *Advanced Methods for high resolution SAR information extraction: data and user-driven evaluation approaches for Image Information Mining*, Doctoral Thesis, Télécom ParisTech, 2011.

[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*, TGARS, 2003.

[3] B. J. Manjunath and M. W. Y., *Texture Features for Browsing and Retrieval of Image Data*, TPAMI, 1996.

[4] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikäinen, X. Chen and W. Gao, *WLD: A Robust Local Image Descriptor*, TPAMI, 2010.

[5] M. Keil, R. Kiefl and G. Strunz, *CORINE Land Cover 2000 - Germany. Final Report*, DLR, 2005.

[6] "Urban Atlas," [Online]. Available: http://www.eea.europa.eu/data-and-maps/data/urban-atlas. [Accessed 16 09 2013].

[7] S. Buckreuss, W. Balzer, P. Mühlbauer, R. Werninghaus and W. Pitz, *The TerraSAR-X satellite project*, IGARSS, 2003.

[8] Fritz, T.; Eineder, M., "TerraSAR-X Ground Segment - Basic Product Specification Document," DLR, Oberpfaffenhofen, 2010.

[9] Werum, "Data Information and Management System for Earth Observation," [Online]. Available: http://www.werum.de/en/mdm/prod/dims/index.jsp. [Accessed 2013].

[10] S. Kiemle, E. Mikusch, C. Bilinski, B. Buckl, D. Dietrich, S. Kröger, C. Reck, A.-K. Schröder-Lanz and M. Wolfmüller, *Data Information and Management System for the DFD Multi-Mission Earth Observation Data,* PV, 2005.

[11] "Service-oriented architecture," [Online]. Available: http://en.wikipedia.org/wiki/Service-oriented_architecture. [Accessed 20 08 2013].

[12] M. Wolfmüller, D. Dietrich, E. Sireteanu, S. Kiemle, E. Mikusch and M. Böttcher, *Dataflow and Workflow Organization — The Data Management for the TerraSAR-X Payload Ground Segment,* TGRS, 2008.

[13] C. Reck, E. Mikusch, S. Kiemle, M. Wolfmüller and M. Böttcher, *Operating tool for a distributed data and information management system,* DASIA, 2002.

[14] S. Kiemle, E. Mikusch and M. Göhmann, *The Product Library — A scalable long-term storage repository for earth observation products,* DASIA, 2001.

[15] D. Espinoza-Molina, M. Quartulli and M. Datcu, "Query by Example in Earth-Observation Image Archives using Data," *IGARSS,* Vols. 978-1-4673-1159-5/12, pp. 6035-6038, 2012.

[16] M. Böttcher, R. Reißig, E. Mikusch and C. Reck, *Processing Management Tools for Earth Observation Products at DLR-DFD,* DASIA, 2001.