# NEW CONCEPTS FOR EARTH OBSERVATION DATA CATALOGUES

*Mihai Datcu*[1], *Alain Giros*[2]

[1]German Aerospace Center-DLR Oberpfaffenhofen Germany
[2]CNES Toulouse France

## ABSTRACT

Conventional satellite data ground segments comprise sensor mode planning, data acquisition systems, data ingestion interfaces, processing capabilities, the catalogue of available data, a data archive, and interfaces for queries and data dissemination. In the present architecture of the systems to access satellite image archives, from the perspective of the information access, the catalogue plays the central role. The catalogue stores the meta information necessary to specify a query. The result of a query is given as a list of images and as quick looks. It is the user who has to browse visually over the quick looks to restrict the result of the query according to his interest. In order to have the possibility to query such catalogues by using the image content, we present a new concept based on an interactive learning of the image semantics, as given by the user.

## THE NEED FOR CATALOGUES

Presently we are faced with a double curse : first data volume explosion and second increasing complexity of image structures and details due to the higher sensor resolution :

- Existing archives already store millions of optical, radar, and other types of data sets.
- Each of these data sets is very rich, typically composed of tens million pixels
- New platforms carrying high resolution imaging sensors will acquire even more data
- Interpretation of these data needs complex techniques, thus the amount of used data is much too small compared with the acquired volume of data.

Not only the data volume is very high, also the information content in the signal is very rich.

## CURRENT CATALOGUES

The state-of-the-art systems for data access are commonly based on static information describing each image such as : geographical location, time of acquisition, type of sensor and in few cases spatial indexes of some semantic information which are computed at data ingestion . In figure 1 is presented the process which produces the static catalog index in such systems.

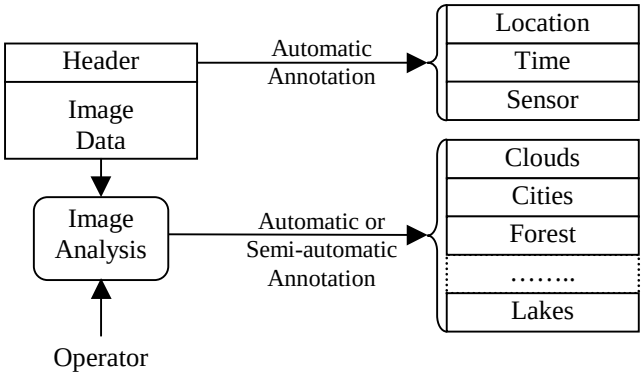


Fig. 1 Generation of the "classical" catalogue entries

As these catalogues are based on fixed meta information structure, the access to the data is limited to the only existing annotation. In Figure 2 is presented the *classical* architecture of an image archive system.

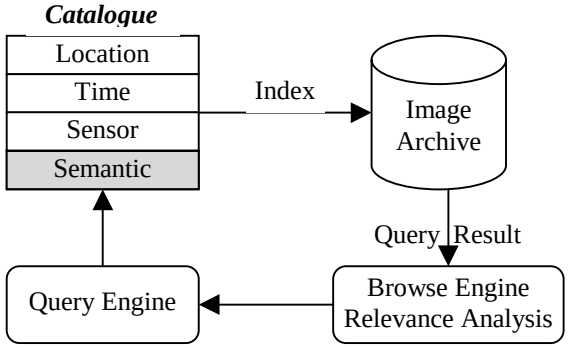


Fig. 2 Use of "Classical" image annotation

Requests can be specified only based on the fixed metainformation catalogue content. The metainformation is a descriptor only of the image data file. The image content information is accessible only for the data sets contained in the results of the query, as quick look images. Their inspection can be done by visual browsing.
As a result of such a concept, and considering the images as sources of information, there is only a very small part of the available information which is used in the querying process. Consequently many images in the archives are not used because they are difficult to be found even though they are relevant.

## NEXT GENERATION OF CATALOGUES

Future ground segments have to be capable of handling multi-sensor and multi-mission data and have to provide easy user access to support the selection of specific data sets, fuse data, visualize draft products and compress data for transmission via Internet. In particular, the search for data sets has to support individual queries by data content and detailed application areas as well as capabilities for automated extraction of relevant features and the application oriented representation of results.

As an effective access to image archives implies a connection with the content of the images itself, we propose a change from "data distribution" to "information distribution" and integration with "interactive value adding". Ideally such systems should exhibit the following functions :

- Tools to enable users to control products before ordering
- Tools to enable users to select data by their information content
- Systems able to provide to the users data and information in an understandable format
- Systems and tools to learn the user conjecture and adapt to user needs

In order to reconcile these complex tasks we developed and demonstrated tools and a scalable system for information retrieval [1,2]. The system is designed to be operated in three steps :

- query by image content from large image archives
- information mining in the set of images obtained as a query response
- scene understanding from the images found to be relevant for the user's application.

With this system, the user can browse the information in an understandable format thanks to an intelligent and friendly interface for data and information query. The system is based on the concepts of information mining and can extract the image information content [6,7]. It also adapts itself in an automatic manner to the users needs delivering a large range of value added products. The architecture of the system is presented in Figure 3. The key components are *the image content catalogue*, the *interactive learning* and *probabilistic query* modules. The system enables the *exploration of the image content catalogue*, as an abstract vocabulary, thus, by an *interactive learning* process, enabling the user to define at semantic level the target of his search. The archive catalogue is dynamically upgraded, thus capturing the users interest.

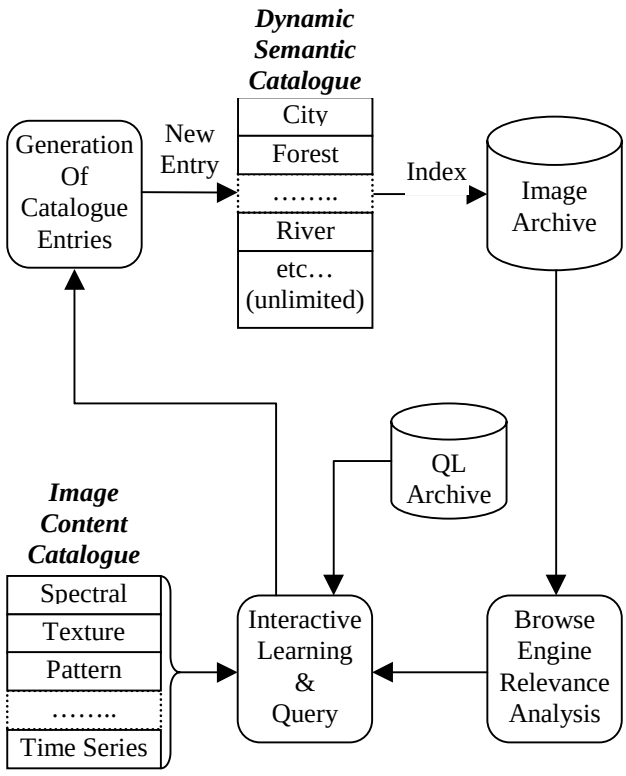


Fig. 3 Information Mining and generation of the dynamic catalogue

One of the main problem to be solved is the production of the *image content catalogue*. In order to keep all the relevant information delivered by the images, it must contain a set of several features such as spectral response, texture properties, geometric structures, … all of these at different scales. The choice of these features is dependent

on the type of imagery and must be carefully studied in each case. The creation of the entries occurs at data ingestion in the archive, as shown in Figure 4.

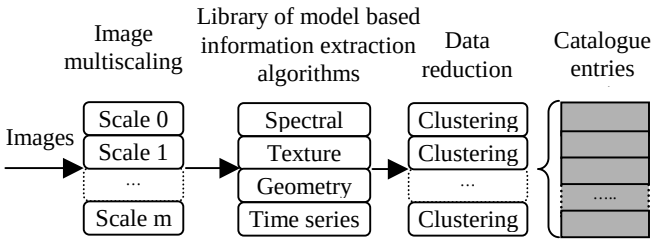


Fig. 4 Generation of the "image contents catalogue" entries

This new structure of catalogue enables the ground segment systems to provide users with new functions :

- Content based image retrieval
- Information dissemination in addition to the state-of-the-art data distribution
- Adaptive systems considering user needs and intelligently assisting the operation
- Interactive pre interpretation of data sets and automatic proposal of relevant information and selected data
- A user centered concept, the system being able to learn and to adapt to the users interests.

## THE IMAGE INFORMATION MINING SYSTEM

Drawing the connection from the image signal to the image content is the key step in providing an intelligent query from large image data bases, thus making the user able to see into the data base before actually retrieving data.

The design concept for query by image content in large image archives aggregates three components [3,4]:

- image content representation and definition of similarity functions
- a multiscale approach with two goals, retrieval at different image scales and fast progressive discarding of non-interesting images
- automatic re-clustering of image archive according to the user query

The elaboration of the system is based on a Bayesian approach for the hierarchic information representation in large image archives:

- image signal modelling and feature extraction
- hierarchic clustering in the joint model " feature space of the image archive
- semantic analysis of the query and user conjecture modelling

The method aims at obtaining a better selectivity of the query process for different subjective user requests, and it is a basis for the design of optimal multi-dimensional signals representation in large over-the-net distributed archives.

At ingestion in the physical archive the image content is extracted and presented in form of signal features and stored in the inventory. The indexing mechanism of the data base management system allows to cluster together all images containing similar content. The similarity is defined adaptively depending on the user conjecture, i.e. the hypothesis used in a certain application [5].

An on line demo system, developed in joint collaboration of the Swiss Federal Institute of Technology, ETH Zurich, and DLR Oberpfaffenhofen, is available on http://isis.dlr.de/mining.

In Figure 4 is presented the result of an interactive learning session, the user is searching man-made structures in an archive of aerial images. In Figure 5 is presented the result of the query.

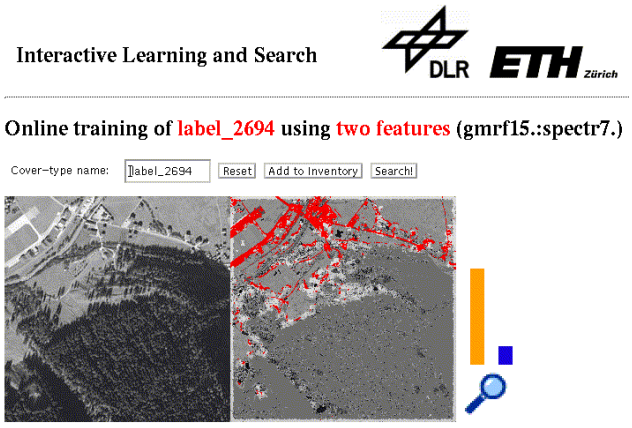


Fig. 4 Interactive learning session for definition of the search target *man-made structures*.

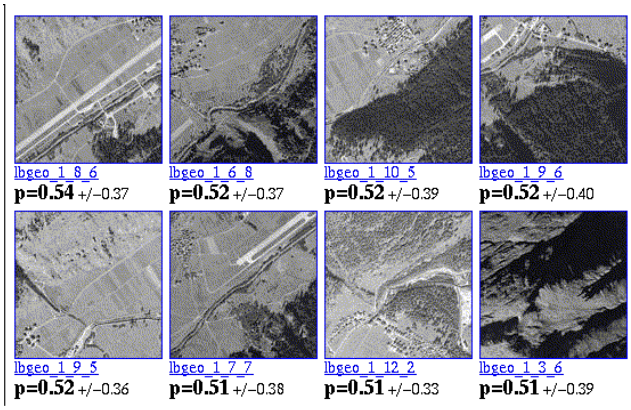


Fig. 5 The results of the probabilistic search for images containing man-made structures.

## CONCLUSIONS

We based and developed a new concept for generation of catalogue systems for large image archives. The new concept enables semi-reflexive self-explanatory capabilities using a natural man-machine dialog. It enables to equip image archives with image information mining function. We regard the mining process as a communication task, from a user centered perspective. The hierarchy of information representation, in conjuction with the quasi-complete image content description, enables implementation of a large variety of mining functions. The concept was demonstrated for a variety of Earth Observation data, further work is done for the development of intelligent satellite ground segment systems, and value adding tools. However its potential is broader, other fields of applications are possible, such as medical imagery, biometrics, etc.

The proposed concept is far from being fully exploited. Presently ongoing theoretical development is going thoroughly into the problematic of complexity of very large image archives. In the case of high heterogeneity observations the complexity and the curse of dimensionality are two key issues which can hinder the interpretation. Therefore, as an alternative solution to the "interpretation", we proposed an exploratory methodology approached from a information theoretical perspective in a Bayesian frame.

## ACKNOWLEDGEMENT

The project has been supported by the Swiss Federal Institute of Technology ETH Zurich, and the German Aerospace Center DLR Oberpfaffenhofen. The authors would like to thank Dr. Klaus Seidel at ETH Zurich for the support to materialize the theoretical concepts.
Presently the concept is under evaluation and further development in the joint collaboration of DLR and CNES.

## REFERENCES

[1] M. Datcu, K.Seidel, S. D'Elia, P. G. Marchetti, 2002, Knowledge-driven Information-Mining in remote sensing image archives, ESA Bulletin, in print.

[2] M. Datcu, K.Seidel, G. Schwarz, 1999, Elaboration of advanced tools for information retrieval and the design of a new generation of remote sensing ground segment systems, in I. Kanellopoulos, editor, Machine Vision in Remote Sensing, Springer, pp. 199-212.

[3] M. Datcu, K. Seidel, 1999, Bayesian methods: applications in information aggregation and data mining. International Archives of Photogrammetry and Remote Sensing, Vol. 32, Part 7-4-3 W6, pp. 68-73.

[4] M. Schröder, H. Rehrauer, K. Seidel, M. Datcu, 2000, Interactive learning and probabilistic retrieval in remote sensing image archives, IEEE Trans. on Geoscience and Remote Sensing, Vol. 38, pp. 2288-2298.

[5] M. Datcu, K. Seidel, 2002, An innovative concept for image information mining, KDD2002, submitted.

[6] C. R. Veltkamp, H. Burkhardt, H.-P. Kriegel (eds.). 2001, State-of-the-Art in Content-Based Image and Video Retrieval. Kluwer.

[7] Ji Zhang, Wynne Hsu, Mong Li Lee, 2001, Image Mining: Issues, Frameworks and Techniques, in Proceedings of the Second International Workshop on Multimedia Data Mining (MDM/KDD 2001), San Francisco, CA, USA, August, 2001.