
skytek
daysha
consulting
VTT
esa
Big Data Mining - Talk #3
'A Collaborative Research Platform to automate Data Mining and Information Extraction experiments'
Paul Kiernan - Skytek
www.skytek.com

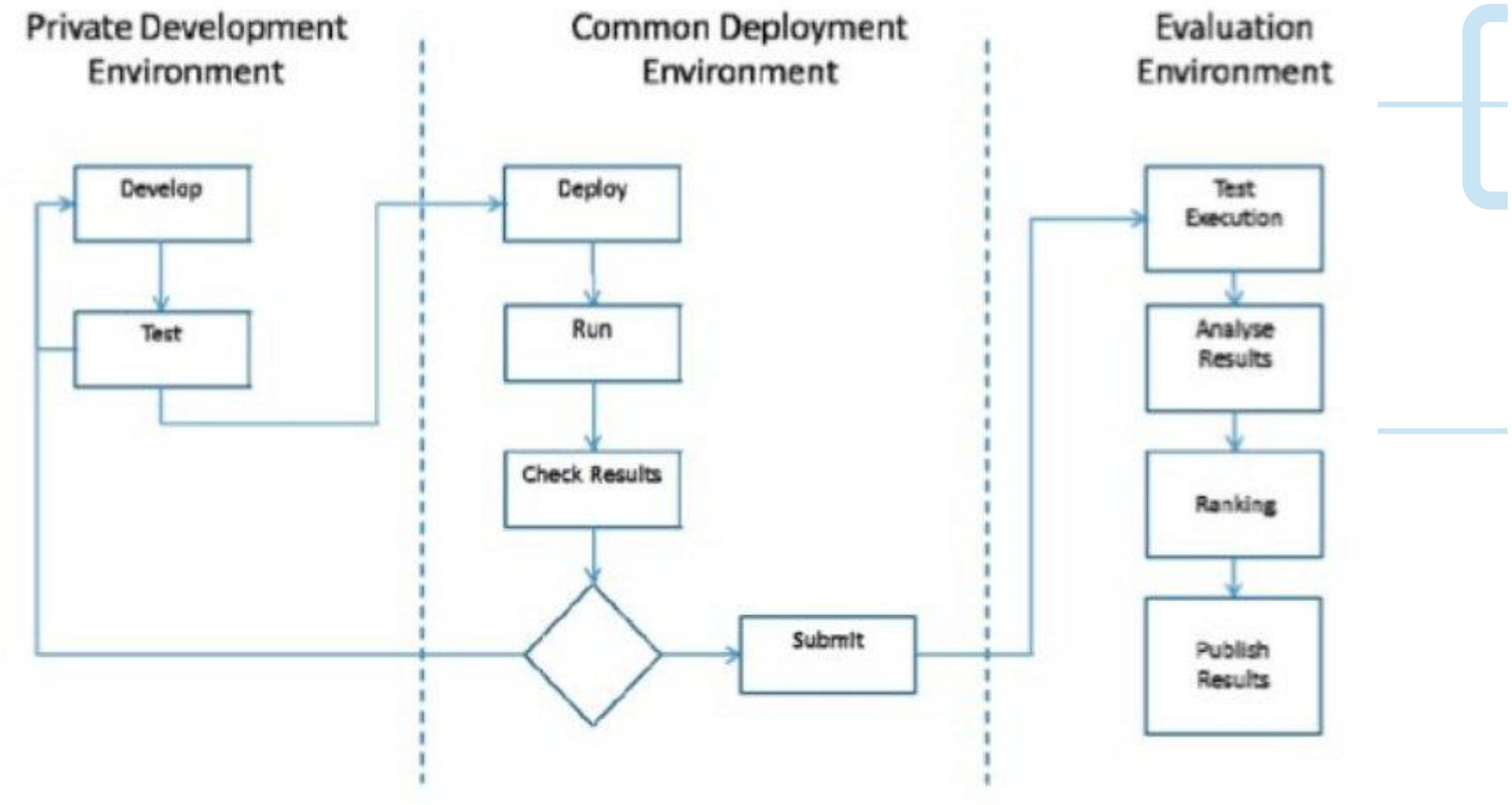
Introduction

- ESA have identified that an impediment to the exploitation of Earth Observation are the current resource and time intensive nature for:
  - Data mining
  - Information extraction
- In this talk we will present the technical architecture for a Common Collaborative Research Platform that will enable the automation of both data mining and information extraction experiments.
- It is expected this platform will be used by members of the space community to investigate problems in a “parallel way” to provide the best solution.
- The project will promote the use of the platform through an ESA sponsored algorithm development contest for Earth Observation researchers.

Core Platform Components

- The primary objectives were identified through user workshops and interviews as:
  - To support the reproducibility and sharing of experimental results
  - To allow contestants to focus their efforts on algorithm development as opposed to environment set up or data preparation
  - To eliminate environment costs associated with participating in the contest
  - To support the contest owner in the set up and operation of the contest
  - Seamless integration into a cloud provisioning environment for access to development and execution resources

Common Collaboration Workflow



Common Environment supported workflow

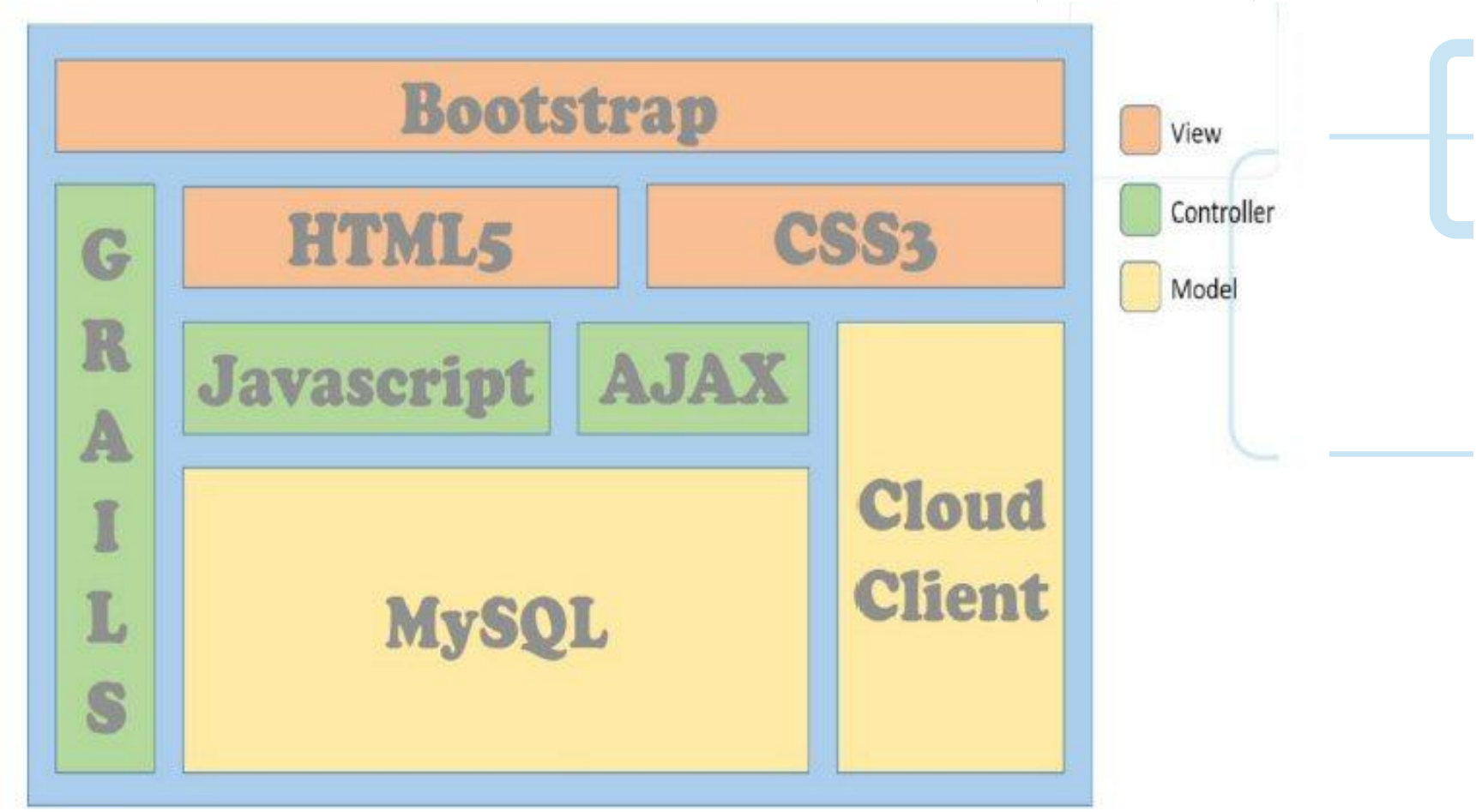
Key Support Functions

- Notification mechanism from contest owner to contestants and evaluators
- Private feeds to contest owners
- Social media discussions with fellow contestants
- Cloud resource environment management, tailored virtual machines for particular contests
- Central data resources and document repository for contestants
- Contest metrics dashboard, time left in contest, number of registered contestants
- Web based easy access to cloud environments
- Evaluators interface for review, evaluation and contest ranking

Platform Architecture

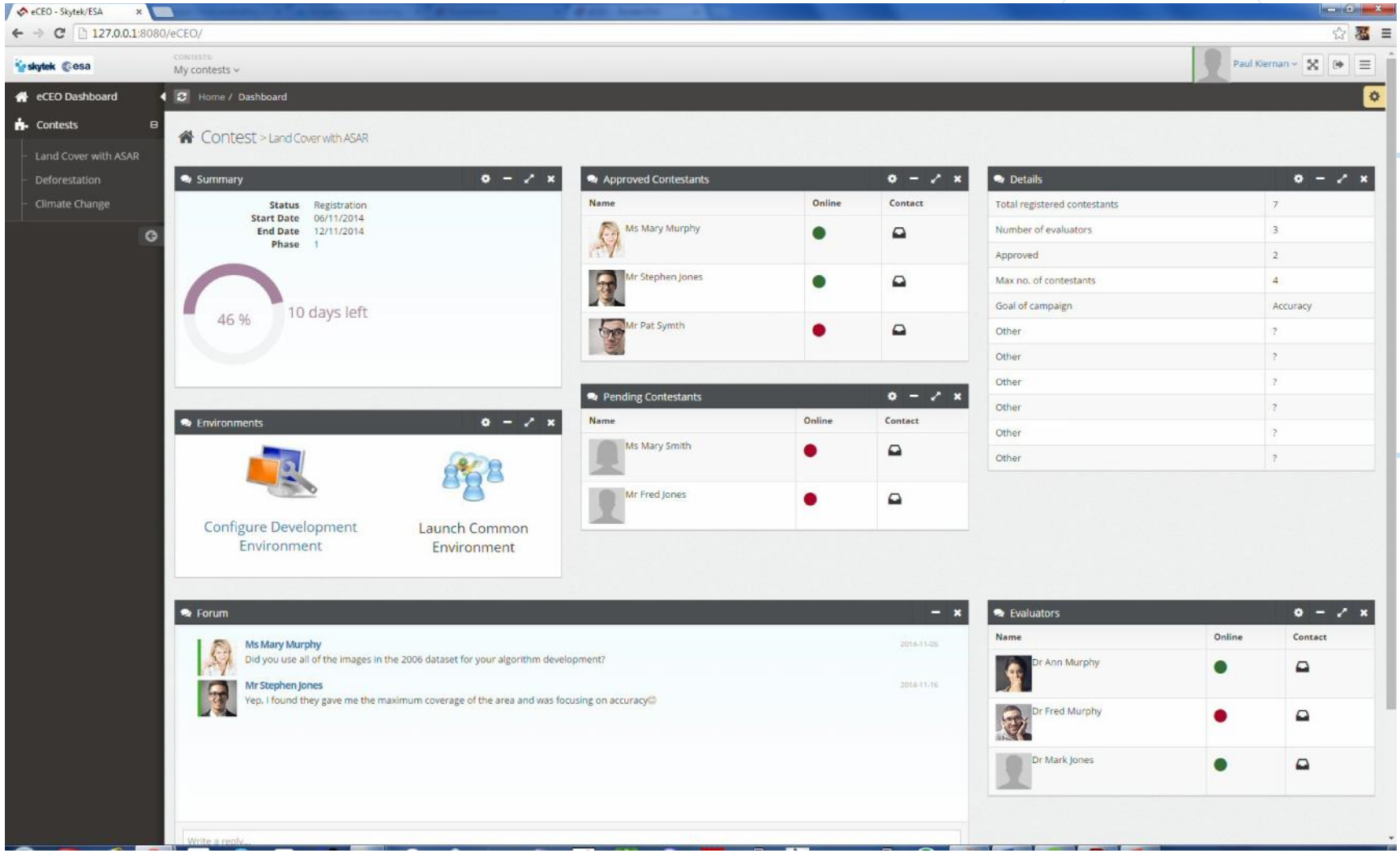
- Platform architecture is a web application system based on a model view controller/services architecture.
- Interfaces with an external cloud computing provider which hosts the private, common, and evaluation environments.
- Selected ‘Interoute’ as the cloud provider
  - Full standards compliant capability virtual machine management
  - Automated VDC API interface.
  - Pan European IAAS
- System uses the standard compliant ‘Apache jclouds’ providing full control of the cloud specific features.
- Create a custom Domain Specific Language (DSL) and Object-relational mapping (ORM) for system.

Platform Architecture



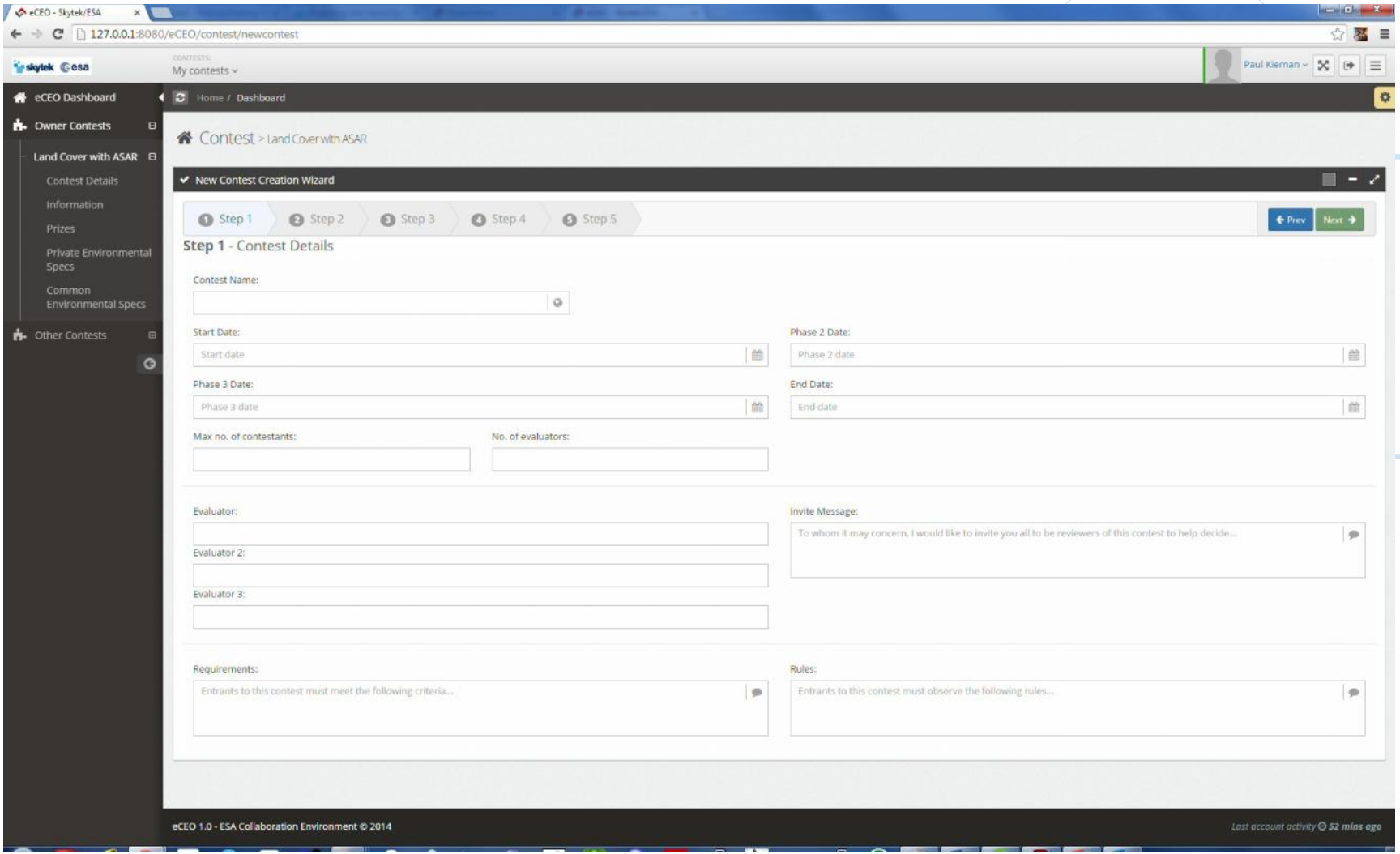
Web application, MVC architecture

eCEO System GUI



Dashboard access to contest information

eCEO System GUI



Wizard based cloud resource/contest creation

System Deployment

- Core system deployed and managed on a Linux based web application server
- Integrates to remote Interoute VDC cloud infrastructure through the jclouds API
- Provides single sign on (SSO) capability for researchers/personnel with ESA central accounts
- Single point of access for personnel to access all their contests, cloud and algorithm execution provided resources
- Leading edge web client technology allowed direct access from a range of web browser.
- Dynamic updates of web client provided through AJAX for updates on metrics and available contestants for live interaction and discussions.

Contest Deployment

- To promote the use of the platform through an ESA sponsored algorithm development contest for EO researchers.
- Three separate contests were initially outlined:
  - Oil Spill Detection
  - Land Cover (selected as initial contest)
  - Deforestation
- Land Cover contest is the general classification and change monitoring of a time series of ENVISAT ASAR WS data for two distinct periods. The competitors will be required to generate land cover maps according to the hierarchical standard defined by the CORINE Land Cover (CLC).
- Contest to be promoted and launched early 2015.

Contest Deployment

- Land Cover contest details:
  - Area of Study, Northern Europe (center Lat/Lon: 49°18’16” N, 6°15’3” E; maximum dimension equal to the swath of ENVISAT ASAR WS which is 405 km)
  - Reference Dates, 2006 and 2009 through optical satellite images on which the CORINE Land Cover classification is based and by the LUCAS information (2009)
  - Contest data set, ENVISAT ASAR WS raw data (405 km swath, 150 m spatial resolution). Amount of data 100 images (approx. 15GB)
  - Preprocessing of data to be done by contestants. Pre-processing will include calibration, georeferencing, despeckle or multilooking and co-registration of multi-temporal SAR data
  - Evaluation, the accuracy of a map will be evaluated through a confusion matrix

Conclusion

- System provides each for EO researchers access to cloud platform resources for algorithm development
- Framework for promotion of research activities of interest to ESA
- Infrastructure and framework for management of users, operators for workflow associated with development, comparison and review of new algorithm development
- Usage of latest GUI methodologies, security provision, developments in open standards and interfaces to automate management of cloud environments for the specific needs of ESA