

# **WASTAC Metadata Archive Proposal**

## **Authors**

Nicholas Bower (nick@nickbower.com)  
Huw Lynch (huw.lynch@optekconsulting.com)

22 November 2006

## ***Executive Summary***

The WASTAC satellite data archive, currently approximately 7Tb and growing at 1Tb per year, suffers from accessibility problems which greatly limit its practical utility. Searching for and obtaining data from the archive is a tedious and labour intensive process due to the fact there is no comprehensive catalogue of data or metadata detailing the characteristics of the individual datasets within it.

This document outlines a proposal to develop a meta-data database to address these issues via an easily accessible and searchable on-line database for use by WASTAC members.

The proposed system is envisaged to be hosted by and make extensive use of storage and computation facilities of iVec. The system shall be coupled to the operational ingest processes which current operate on iVec infrastructure.

## Overview

The WASTAC satellite data archive, currently approximately 7Tb and growing at 1Tb per year, suffers from accessibility problems which greatly limit its practical utility. Searching for and obtaining data from the archive is a tedious and labour intensive process due to the fact there is no comprehensive catalogue of data or metadata detailing the characteristics of the individual datasets within it.

This document outlines a proposal to develop a meta-data database which aims to solve these issues by;

- cataloguing scene meta-data in a relational database
- cataloguing reduced resolution observations and/or derived parameters (i.e. cloud, sun-glint) in an on-line spatial-enabled database
- allow searching of the above database data to identify scenes of interest in short period of time
- provide a web based portal interface to the search facilities by which registered users can search for and download data from the WASTAC archive.

The above points list our core objectives for this work. While investigating this project we found other potential objectives which are not yet considered essential;

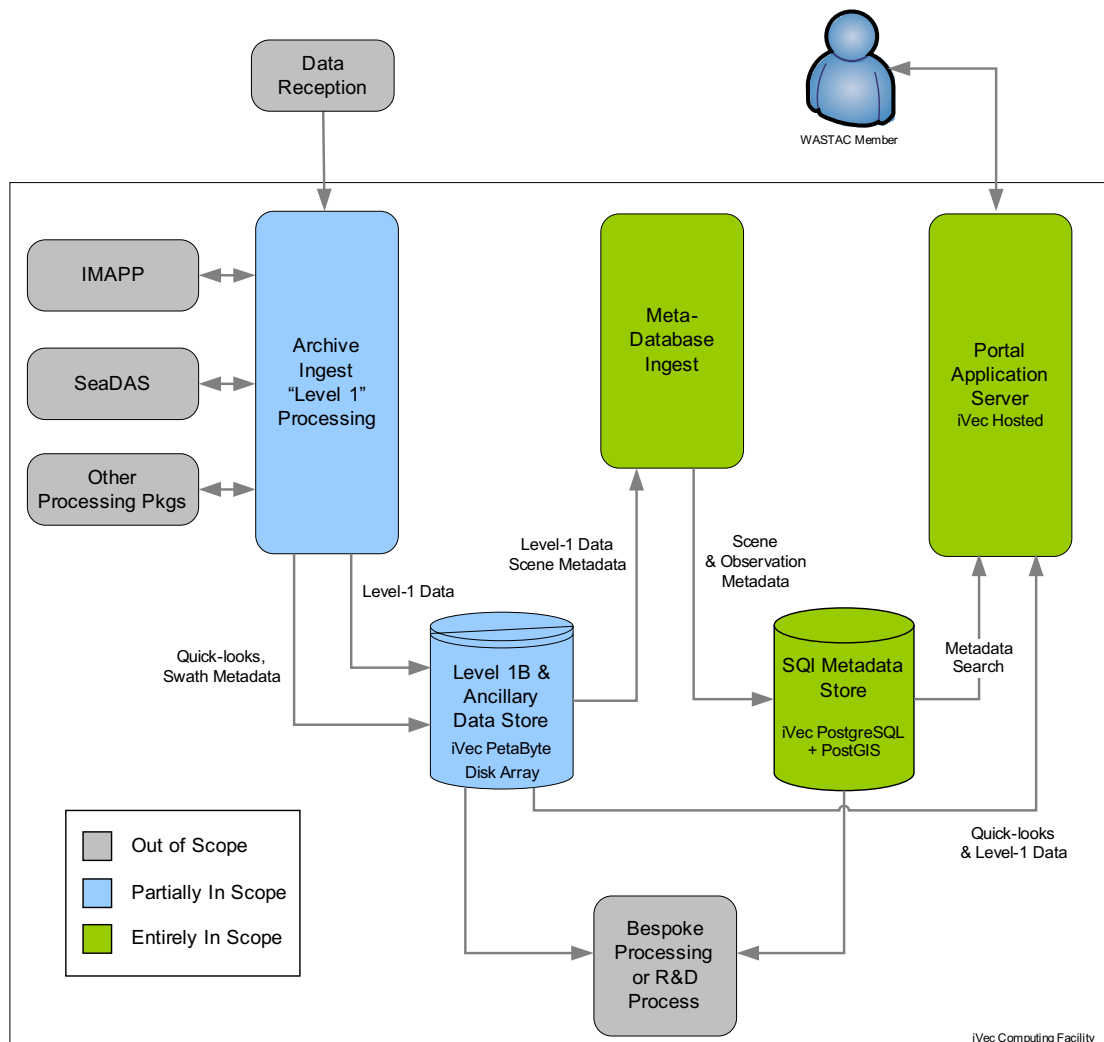
- on-demand processing of level-1 data sets to produce specific well-understood and accepted products (such as sea surface temperature for the fishing industry)
- conversion to delivery of data sets to users in application-oriented formats (such as SST in GeoTIFF format projected into UTM)
- data rights enforcement; parts of the data in the WASTAC archive is rights free, however some (such as SeaWiFS) can only be distributed under license. A portal that provides access to known authenticated users must comply with data licensing restrictions.

The proposed system is envisaged to be hosted by and make great use of storage and computation facilities of iVec. The system shall be coupled to the operational ingest processes which current operate on iVec infrastructure.

The WASTAC archive contains data from a long line of instruments over many years. This proposal aims to produce a metadata framework that can be back-filled with the full set of historical data, but does not aim to complete that task in full. Initially we aim to integrate a selected set of current sensors, which will be identified during the early stages of the project.

# Functional Overview

The diagram below presents a high level functional view of the system and related WASTAC processes. Colour-coding indicates which aspects are in and out of scope of the proposed system.



**Entirely In Scope** Developing these processes is part of the proposed work

**Out Of Scope** These processes may already exist and are completely out of scope of the proposed work

**Partially In Scope** Processes that exist and are closely related to the proposed work. Some degree of interoperability is essential, so some changes may be needed for these systems.

# Metadata

## Scene Metadata

Each satellite scene has obvious attributes that are candidates for being treated as metadata within the system. This data is traditional relational data and is highly suited to storage in a relational database. Attributes are likely to include satellite, sensor, start/end time, scene bounding shape, processing software version, day/night, processing date among others.

## Observation Metadata

It is not feasible to search all observation data in the WASTAC archive using readily available technology and development of such technology is considered overly ambitious. It is expected that attempting to search the full observation dataset will remain infeasible over time, because the volume of data being generated by future satellite instruments will increase at least as fast as the capabilities of commodity storage and search technology.

To facilitate near real-time search capability over large quantities of satellite data, the proposed system will search a reduced version of the archive. Specifically the meta-database will;

- index a selected subset of observation data (ie specific bands) and / or specific derived data (such as cloudiness or the presence of sun-glint) that are deemed useful in identifying interesting datasets or excluding in appropriate data sets.
- search a spatially reduced version of observation data to achieve a balance between data volume and storage/search feasibility

The proposed means of data reduction is to establish a fixed grid over the Earth (or area of the Earth, such as all of Australia) and store summarised observation data for each of those grid cells. Initial discussions have indicated that a 10km x 10km grid would be sufficient resolution for search requirements while not resulting in an impractical storage volume.

The following diagrams illustrate an arbitrary 10km grid over Perth and Shark Bay.

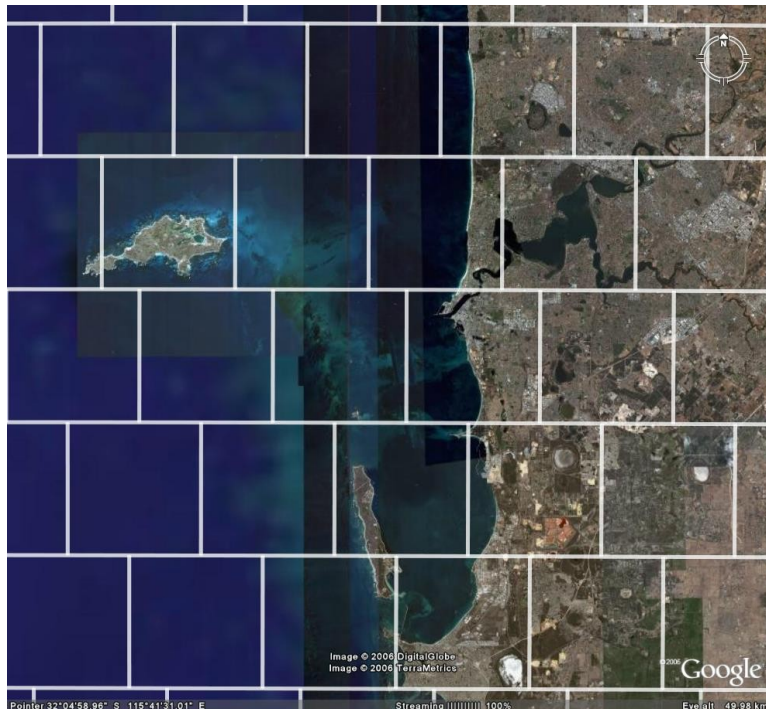


Figure 1 – 10km grid over Perth, Western Australia

A 10km spatial grid may not be ideal – the early stages of this project will aim to identify a suitable cell size. Over time the preferred cell size is likely to change – either to increase search accuracy, or simply to take advantage of increased database capabilities.

To accommodate cell-size changes, the system will be specifically designed to allow re-calculation of the observation component of meta-database from the archive data. For this to be practical that task will have to be a highly automated batch process which could be invoked on an infrequent as needed basis.

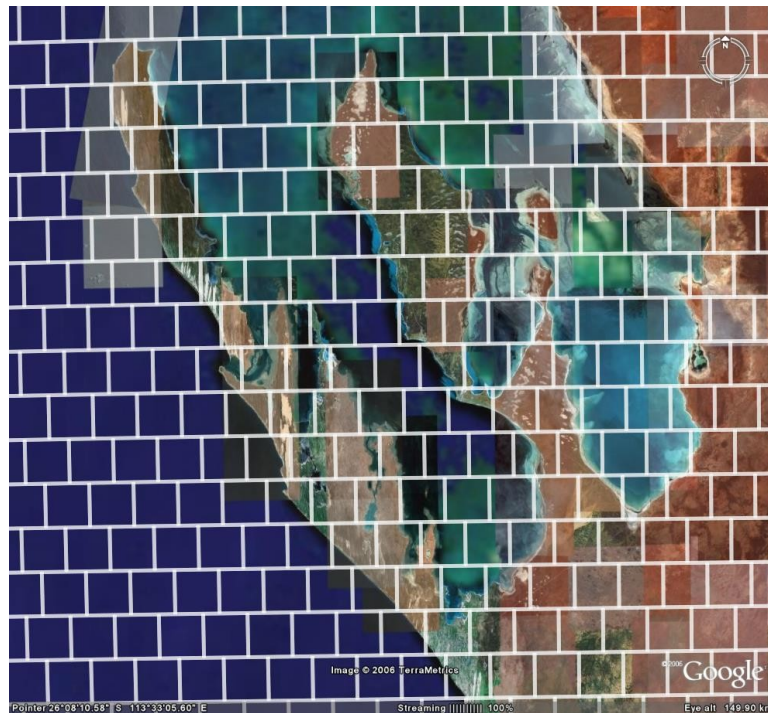


Figure 2 – 10km grid over Shark Bay, Western Australia

## **Proposed Project Plan**

The project is conveniently broken into incremental phases. The suggested development approach is to prototype the system and lead the prototype into production form via refinements.

### ***Project Phase 1 – Requirements and Feasibility***

#### **Objective**

**Investigate and fix key aspects and requirements of the system, determine feasibility of storage and processing strategy of the system.**

#### **Deliverables**

*Essentially Documentation Covering*

- Scene meta-data entities
- Observation meta-data entities
- Initial meta-database relational schema
- Observation data reduction methodology
- Determine archive file and volume structure
- Prototype metadata generation, observation reduction scheme and database ingest
- Estimate metadata storage requirements per-scene, per-year and processing times
- Refine interface between ingest and metadata components
- Identify for which sensors the archive will generate metadata

#### **Required Participants**

- iVec infrastructure support
- WASTAC guidance on metadata, processing systems

## ***Project Phase 2 – Core Objective Implementation***

### **Objective**

**Deliver operational metadata database integrated with pre-existing automated ingest system and simple web interface to metadatabase.**

### **Deliverables**

*Operational software meeting core objectives, installed and operational on iVec infrastructure. To be provided with accompanying operational, user and developer documentation.*

- Establish online data file archive maintained with automated processes
- Users can log into database with SQL client and browse the archive metadata and determine where to locate particular data files in the volume file system.
- Users can search for and download archive data files using a basic web interface
- Production proofing of the system to improve operational reliability
- User, Operational and Developer documentation

### ***Project Phase 3 - Future Options***

This phase is not yet defined. All of the following are only listed for consideration. The exact objectives of this phase will evolve throughout phases 1 and 2.

#### **Possible Objectives**

- Layer-2 and beyond processing on-demand
- Enhanced web portal providing graphical search and stream-lined web experience
- Integration of historical datasets and instruments
- Enhanced data format delivery options (possibly via OpenDAP)
- Integration with iVec APAC grid computing middleware

## Base Software Components

The system will be developed keeping simplicity in mind (to increase transparency and maintainability by non-expert software staff) and make use of free open-source software when possible to minimise development effort. Where ever possible scripting languages will be used in preference to compiled languages. We do not yet envisage a requirement for any commercially licensed software.

The following open source packages are likely to feature heavily in the resulting system;

- PostgreSQL<sup>1</sup> – A full featured open source database
- PostGIS<sup>2</sup> – Spatial extensions to PostgreSQL
- Zope<sup>3</sup> – A full featured application server suitable for portal development
- Python / PERL – General purpose scripting languages
- Linux (iVec operate on SuSE Linux)

Note that the licenses associated with the above software (commonly referred to as “free and open source”) are likely to require that the developed software is also “free and open source”.

---

<sup>1</sup> <http://www.postgresql.org/>

<sup>2</sup> <http://postgis.refrations.net/>

<sup>3</sup> <http://www.zope.org/>

## **Caveats**

### ***Requirements Specification***

We have presented a proposal and associated estimates for works based on the best understanding of the requirements that we have been able to piece together thus far. The exact requirements will vary in stage 1 as a result of interactions with WASTAC and may affect reasonable estimates of stage 2.

### ***Professional Software Developers Acting Independently***

We (individuals proposing to develop this system) are capable software developers with extensive experience in science, software integration, data-processing and web technologies. We also have degree of affinity to the WASTAC cause. However we are NOT a commercial “software development house” as such. In this situation we lack typical benefits / overheads of a software house including but not limited to;

- **longevity** - We do not work independently in this area of business full-time and will most likely not be available to do so in the long-term.
- **liability** - As individuals we do not have a significant corporate infrastructure that can be pursued in the event of perceived damages. Consequently while our affinity with WASTAC ensures our best-efforts, we must aim to avoid or minimise any financial liability in contractual arrangements.

### ***On-Going Costs of Bespoke Software***

In commissioning the development of bespoke software, the funding entities must take into account and budget for the on-going costs of maintaining the resulting products. Software that is developed will by nature of the software industry, reach an operational end-of-life if not maintained or not kept up to date with 3rd party products on which it depends. It is hence critical that this project is not viewed as a fixed or once-off budget item in isolation.