Report on HiRISE Catalog System Review

Introduction

The HiRISE Catalog informal review was conducted at the HiRISE Operations Center (HiROC), University of Arizona, on October 13, 2004. The review team, see table of review team members below, had individuals with expertise in database management systems and computer science. Several are involved in supporting active NASA flight projects.

Reviewer	Affiliation	Project / Department
Kate Crombie	University of Arizona	Odyssey/GRS
Karl Harshman	University of Arizona	Odyssey/GRS & Phoenix/TEGA
Douglas Hughes	Jet Propulsion Laboratory	Project Elements
John Ivens	University of Arizona	Cassini/VIMS
Jascha Sohl-Dickstein	Cornell University	Mars Exploration Rover
Richard Snodgrass	University of Arizona	Computer Science
Alice Stanboli	Jet Propulsion Laboratory	Planetary Data System

The team was asked to ascertain the design of and planning for the HiRISE Catalog System (HiCat) in the context of the HiRISE Operations Center. The review charter was:

- Determine the technical soundness of the HiCat system in terms of
  - Table Structure
  - Security
  - Performance (load-balancing, replication, etc.)
  - Data Integrity
  - Reliability
- Assess the HiCat design and planning against its requirements
- Determine if the implementation plan is adequate

A HiRISE HiCat Review web site, http://hirise.lpl.arizona.edu/HiROC/HiCat/Review, was developed to provide background and presentation material for the review team. This web site will also house follow-on information, review results, and intended action to be taken by the HiROC team to address issues identified by the reviewers.

During the first part of the review, HiROC staff presented material about the HiCat functional requirements, design, and implementation plan. The second part of the review provided an opportunity for open discussions with the review team. The review team was asked to identify problems and recommend solutions for HiCat concerns. The review team filled out Request for Action (RFA) forms capturing issues and

concerns and offering recommendations on how these could be addressed by the HiROC staff.

Review Summary

A general summary of reviewer comments and suggestions is provided below. RFAs developed by the review team follow the review summary.

Alice Stanboli

- Design generally solid
- Develop separate catalog views to keep operations hidden from public
- Some tables have too many fields. Separate tables with joins may be more efficient and faster, 30-40 columns are generally a good limit.
- A view that covers multiple tables may offer improvements in performance.
- Consider adding gap maps to the database

HiROC team will contact Alice on what fields PDS needs for data product deliveries and what fields are typically queried many times.

Richard Snodgrass

- Likes the database in the center of HiROC and overall structure of the project
- Vetting image suggestions by the science team will be a bottleneck. Consider what information is really needed. Suggest simple and advanced versions of HiWeb.
- More work needed on science themes needing multiple images (not allocation) to do science.
- Provide a numerical value for criticality of each parameter of an image suggestion.
- Suggest HiROC team look into software for automatically reviewing conference paper submissions to get ideas on approach.
- As part of EPO, provide web resource describing step-by-step image processing procedure.
- "Condor" may possibly help with processing pipelines.
- Two conflicting requirements (Public access / Operations) - provide two systems supporting DBMS activities--one inside and the other outside the firewall
- Track changes to pipeline processing. Develop a mechanism that can answer the question ten years hence: "what did we do in the processing?"
- PDS will permanently archive the data, but information will be lost about processing unless a mechanism is put into place. Work with PDS on processing history information to be passed to PDS.
- Entries in the database should be time stamped.
- Expand use of user scenarios and story boards.
- Integrity checks are good but needs to be done at time of insertion. "Trigger" tests on inserts may be useful.
- HiCat is complex enough to do a formal review of the tables.
- Check out web caching expertise on campus.

Douglas Hughes

- Stresses importance of using web caching on DMZ for critical times of the mission (press conferences, start of mission, etc.)
- NASA has a contract with "utouch", possible coordination?
- Web caching service takes the heat off of HiWeb
- Configure manage HiWeb to coordinate changes with HiCat. Needs "Rollout coordination"
- Cache top 10 images and/or 10 latest images.
- Develop an automated data recovery system and method to bring up another server while the current server is running.
- Develop an operational big picture to make sure no one is left out in the cold
- JPL has tools to test web servers with external loads. More than just performance testing but also determines what breaks.

Jascha Sohl-Dickstein

- Add a products table for raw products downloaded from RSDS.
- True/False for data quality is not sufficient. Provide test field to describe problems, possibly a numerical value to quantify. Validators need to input this manually.
- Consider flagging EDRs that are not go into RDR processing.
- Add data quality table that applies to both EDRs and RDRs
- Provide tracking of versions, including planning of observations.
- Link observations by hypothesis as opposed to just a science theme.
- Mapping across instruments is an issue, strong procedure needed for coordinated observations
- Warning FEI can fail silently. Crosscheck download with other tools.
- Science rationales need to be broken into pieces - easier to search for keywords.

John Ivens

- Data product headers should include all the steps in their production
- More effort needed on how desires of science team are incorporated into the planning
- MySQL should standup to HiROC use
- Optimize by developing a replicated server for fast read queries (no inserts)

Karl Harshman

- Noble goal to make data available to all ASAP
- Version tracking very important but missing in HiCat
- Address backup system to cover file system, not just catalog
- Need a backup of log files
- RDR files need to have information on calibration files, SPICE versions, EDR versions, and processing.
- Provide a checksum as method to see if untracked changes occur (errors can occur by transferring data).

Kate Crombie

- Issues with data validation: develop a Standard Operating Procedure (check list)
- Develop a plan on handling backlogs
- Develop feedback loops between validator and targeting team
- Define mechanism for data validation
- Version control is missing in HiCat design.

#	Problem / Issue	Recommended Action	Category
Joscha Sohl-Dickstein			
1	Difficult to reproduce old RDRs as software and calibration parameters evolve.	<> Have the EDR level produce an RDR table row with newest parameters, and marked as RDR not yet created. <> Have the RDR generation software look for such rows in the RDR table, and then create RDRs based on the parameters in the RDR table <> This will force more stringent versioning.	Reliability
2	Might consider tracking statistics to track data product sizes.		Data Integrity
3	Difficult to keep track of which observations are linked/dependent on each other	Allow users to create hypotheses and associate proposed observations with them. (in addition to the very bread science themes, they are currently able to associated observations with new hypotheses table-linked to users, themes, and observations.	HiCat
4	Plan to track/relate data products and observations across MRO instruments. Not yet well developed		HiRISE s/w
5	Validity/quality has no room for shades of gray (nothing but true/false fields)	<> Create new data product or observation level quality table. <> Include a text field for validator comments. <> Include a binary and text field dealing with resolution problems. <> Fields for dropouts, saturation, etc. Perhaps a lookup table allowing newly discovered reoccurring problems to be added.	Data Validation / Tracking
6	Tables (i.e. EDR) may be large/unwieldy	Consider breaking into smaller tables, by subtopic (header information, tracking information, quality information...).	HiCat
7	There is no DB visibility between creating a command text for uplink and creating an EDR after downlink	<> Create a "raw data products" table and populate it as soon as a product is received from JPL (before EDRs are created). <> Consider tracking uplink radiation logs onboard file system	Data Validation / Tracking
8	FEI unreliable with occasional silent failure	<> Automatically compare planned observations against raw data products (this will also catch non-FEI problems between when the observation leaves your grasp and when it comes back). <> Alternatively, periodically run a script at JPL that polls the DB and compares the raw data products it knows about with what's in the file system at JPL.	HiRISE s/w

#	Problem / Issue	Recommended Action	Category
Kate Crombie			
9	Data validation has a number of issues to be resolved : 1) How long will it take for a validation to get through each image? Will there be a Standard Operating Procedure (checklist) 2) How will back logs be dealt with? 3)What are the feedback loops between validator and the targeting people? 4) What is the mechanism of validation (web, java?)		HiRISE s/w
10	Version control on the RDR Data Products. It may be helpful to add version numbers to RDR products to track how many times with what software/calibrations an RDR has been produced. Also, with any data re-release to the PDS, the release number is incremented. This may be a way to track the versions.	At a minimum, the processing stream elements should be recorded in the image header/metadata.	Reliability
Karl Harshman			
11	Reduced products may not have the information about how they were produced and what SPICE kernels they were produced with.	Keep a record of the information of how a product was produced with the product.	HiRISE s/w
12	Backup of only part of the data	Since DB is primarily keeping paths to files instead of the actual data, all these files should be backed up.	Backups
John Ivens			
13	Header information needs to contain all steps, versions of files used, etc. in the file header itself. Especially because you don't store different file versions. History section of the PDS.	Do it. All parameters, files, etc. SPICE files, hot pixel maps, etc.	HiRISE s/w
14	Planning - Do you intend to work together with all teams and share planning information or do you want to map out segments of tour where you have primary targets and riders?	Impacts the planning software - how to handle the data volume requested by each team and how to prioritize intra-team requests. Power issues, etc. may be affected.	HiRISE s/w

#	Problem / Issue	Recommended Action	Category
Douglas Hughes			
15	Without using an external web-based testing suite, certain problems can't be flushed out, including load and performance	Consider, within budget, at least two large test suites for web testing - "Mercury Interactive" (set of commercial tools)	Performance
16	HiWeb Bandwidth Caching & DMZ: A significant public engagement will crush UA net infrastructure.	Place HiWEB caching engines in a DMZ. Consider a caching service for the first few weeks of the science mission.	HiWeb
17	Given the probable lifetime of the mission and criticality of the web I/F, it seems that the informality of the relationship with NASA ARC developers and the lifecycle has some risk.	Formalize HiWeb development process. Consider a stable home for the long-term maintenance, complete with a reasonable lifecycle.	HiWeb
18	HiWeb: It seems that each HiWeb interaction causes load on production database.	Cache "top 10" favorites, cache "newest Images"	HiWeb
19	HiCat Database backup/recovery: A lack of proven and automated DB recovery is a problem in an operational environment.	Develop the process and scripts to recover the database.	Backups
20	Database change coordination: Lack of a formal process for DB change coordination will hamper operations.	Provide close coordination (process and possibly tool) to facilitate DB changes. Be careful of breaking applications. Check dependencies.	Reliability
Richard Snodgrass			
21	Web access may overwhelm bandwidth	Consider caching web pages; Dr. Bongki Moon at the Computer Science department does research into web caching and has a nice caching system on top of Apache.	HiWeb
22	Data Base design incomplete	Do a formal database design review (review at the table and field level). I would be happy to participate in such a review.	HiCat
23	Integrity checking	Suggest instead implementing this at update time rather then periodically. <> triggers; <> sanity checks.	Data Validation / Tracking
24	What will interactions with scientists look like?	<> User scenarios are a great idea; <> expand into regression tests. <> Also suggest story boarding of user interfaces	Miscellaneous

#	Problem / Issue	Recommended Action	Category
25	Need to do versioning; <> pending observations; <> products. Need to timestamp most data for date provenance	Utilize existing, known temporal database ; <> utilize append-only semantics where appropriate: database data and files. <> utilize a stratified approach for data manipulation; <> utilize RAID for file storage.	Reliability
26	Pipeline composition changes; This is a unified aspect of every product	<> Version the description of the pipeline of the data in the database (as well as in CVS). <> Make this data available to scientist (may be available indirectly). <> How to answer these questions years later.	Reliability
27	How to ensure: <> Processing requirements especially observation scheduling; <> disallowing external intrusions; <> supporting massive browsing by the public	Consider physical partitioning of database, on two different version of the DBMS instance with well-defined data transfer between, and with critical DBMS more closely guarded; Front end versus backend.	Performance
28	Multiple conductors on different machines	Perhaps use "condor" to manage this.	HiRISE s/w
29	Students might be interested in the process by which an image was derived - emphasize image is not raw data but highly processed	<> Make process information graphically visible via HiWeb. <> For a few images show result of each processing step to understand what is going on.	HiWeb
30	<> Vetting of suggestions: - school suggestions very different then science suggestions: 1) Volume, 2) input information, 3) specificity of requirements (e.g. students are not going to understand binning) <> How can multiple images be suggested; <> Can you non spatially oriented requests; - How can ranges of parameters be specified; <> How can urgency of parameter values be specified.	<> Differentiate public and educational user interface vs. scientific suggestions; <> Vetting procedure: - perhaps use conference review metaphor (see conference management software for examples)	HiWeb
Alice Stanboli			
31	Table definitions may contain a large number of columns that may impact performance. This is due to the fact that one record may need to be stored in multiple pages.	Group columns into separate tables and join on integer primary keys. This is often faster on selects.	HiCat
32	Potential impact of remote user access on DB server. This could cause slowdown in pipeline operations.	Decouple access to server. For example, have one dedicated server for operations and another mirror DB server for remote users.	Performance
33	Need to give different views of the meta data to users that have different needs.	Could use MySQL views to translate the metadata. This will also provide a unified way of viewing meta data across applications.	HiCat