Wide-field Infrared Survey Explorer (WISE)

Archive Preparation and Transfer Procedures

Version 1.5

20-January-2010

Prepared by: Lin Yan

Infrared Processing and Analysis Center
California Institute of Technology

WSDC D-C003

Approved By:

Lin Yan, Archive Scientist, WISE Science Data Center

Davy Kirkpatrick, QA Scientist, WISE Science Data Center,

Roc Cutri, WISE Science Data Center Manager

Revision History

Date	Version	Author	Description
04-Sept-2009	Version 0.5	Lin Yan	Initial Draft
26-October-2009	Version 1.0	Lin Yan	
20-January-2010	Version 2.0	Lin Yan	

Table of Contents

1 INTRODUCTION

This document describes the procedure with which the WISE data product will be loaded into the archive database. The mission requirement is that within 1 week of receiving the raw data at IPAC, the WISE archive will deliver the data product, including the source catalogs and reduced images to the science team. Therefore, the WISE archive will need to load new datasets every 3 days, i.e., archive data loading frequency of twice a week.

2 FLOW CHART FOR THE ARCHIVE LOAD AND TRANSFER PROCEDURE

There are two major parts which determine the WISE data being successfully loaded into the database in a timely fashion. First is the data preparation by the WISE team, and the second is the data loading and delivery to the science users by the IRSA team. Here we visually illustrate the archive data loading and transferring procedure with the flow chart below. We summarize the data loading procedures as follows.

1. When the raw data comes down WSDC, it will be ingested and processed through the WISE pipeline.

2. The WISE QA team will examine and validate the processed data. The QA scores will be stored in a database, from which the archive loading script will read the frame indices to determine which scans will be ready and prepared for loading. The QA work is done continuously, and follows every batch of dataset down linked from the telescope and processed by the WISE pipeline.

3. The WISE OPS team will then prepare the source tables and stage the data for IRSA loading by running the archive loading script. Fits images will be copied from the WISE server to the IRSA machine. Specifically,

    a. The script will check the frame indices in the QA database and determine if the data is good for archiving, then assign each dataset with a loading ID. This ID will allow us to keep a track of each dataset loaded into the IRSA database at a specific time. If any dataset has problems and needs to be reloaded, and we can use this loading ID to remove the data without touching rest of the database.
    b. The script will also generate checksums for the fits images. When the IRSA team transfer these data, the checksums will provide an easy and quick way to ensure the data integrity.

c. The archive loading script has two main parts, DBPREP and IMPREP. DBPREP will prepare the tables, attach the loading ID to the data, and stage the tables in a specific disk area from which the IRSA database loading script will fetch the data.
d. IMPREP will prepare the image fits data, and generate a list of files which can be used as an input to an ftp script which will copy the specific set of fits images from the WISE disks to the IRSA data disks.
e. The archive loading script will also make data file manifests which will keep records of files which have been transferred from WISE to IRSA.
f. Besides running the archive loading script, the WISE OPS team will eventually also execute the scripts for the fits image transfer and the IRSA database loading. The details on how this can be done will be sorted out after a few weeks of nominal survey mission.

4. The WISE data will be loaded into the IRSA database. The current plan is following:

   a. The IRSA team has provided a server which deals with only the WISE data. The wise catalog and image archive are both staged on http://hades.ipac.caltech.edu/applications/ Gator (catalog) and http://hades.ipac.caltech.edu/applications/WISE/IM (image archive).

   b. The full source catalog database will be organized into individual segments, and the data from each loading forms one segment. Each segment database will be tagged by a load ID so that in the event of failure, the specific set of data will be traced easily and replaced without interruption of the full database.
   c. The IRSA team will provides scripts for both loading the source tables into the IRSA database as well as transferring the fits images. Currently the loading is done by the IRSA team. We are in the process of discussing a more automated way, and the details will be worked out after a period of initial data archiving.
   d. The script for transferring the fits images will be limited to a reasonable speed, for example, 18MB/sec, in order not to impose any interruption to the WISE pipeline processing.

5. After the new dataset is loaded, the WISE QA team and the external (outside WSDC) WISE science team will be able to view the newly loaded immediately. If no problems have been identified, the newly loaded dataset will be automatically a part of the full archive. If problems are found, the IRSA team will be alerted and the newly loaded dataset will be removed from the access. Both WISE archive and QA team will work together with the IRSA team to identify the nature of the problems and re-load the archive with the correct dataset. Figure 2 shows the flow chart how the WISE QA team will interact with various groups to

validate the newly loaded WISE data and keep track of any anomalies. As stated before, DBPREP and IMPREP will generate two manifests listing all of the files transferred from the WISE to IRSA disks. The WISE QA will perform checksum on the new data to be sure that the correct quantity of data being loaded into the archive.

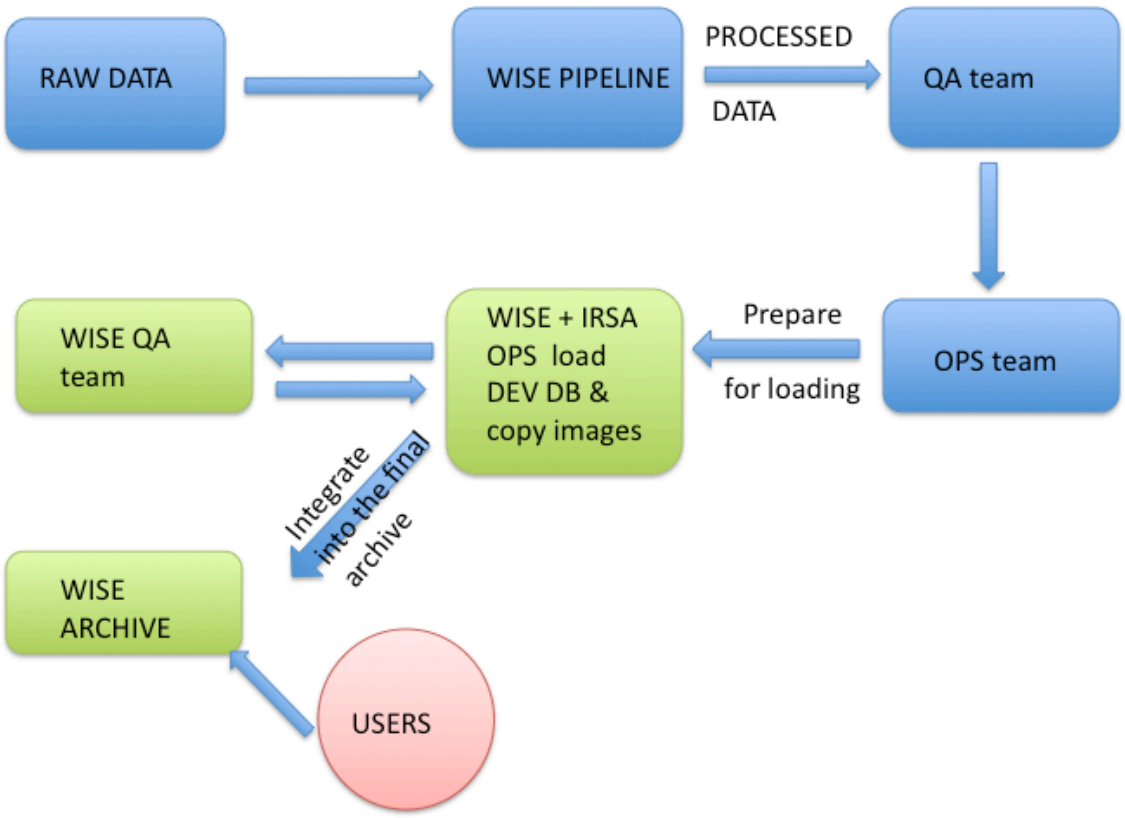


Figure 1. The data flow chart from the raw data to the final processed and archived data during the WISE mission operation.

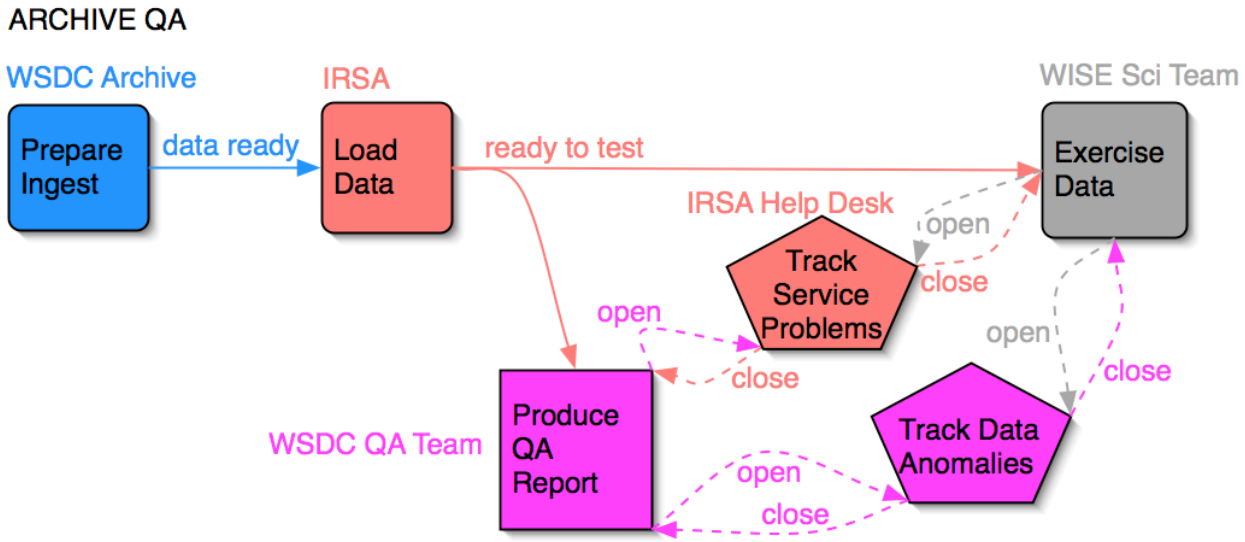


Figure 2. The detailed steps on how the WISE archive quality control and assurance is performed. We note that unless any problem is identified, the newly loaded data will be accessible by the science team as well as the WISE QA team.

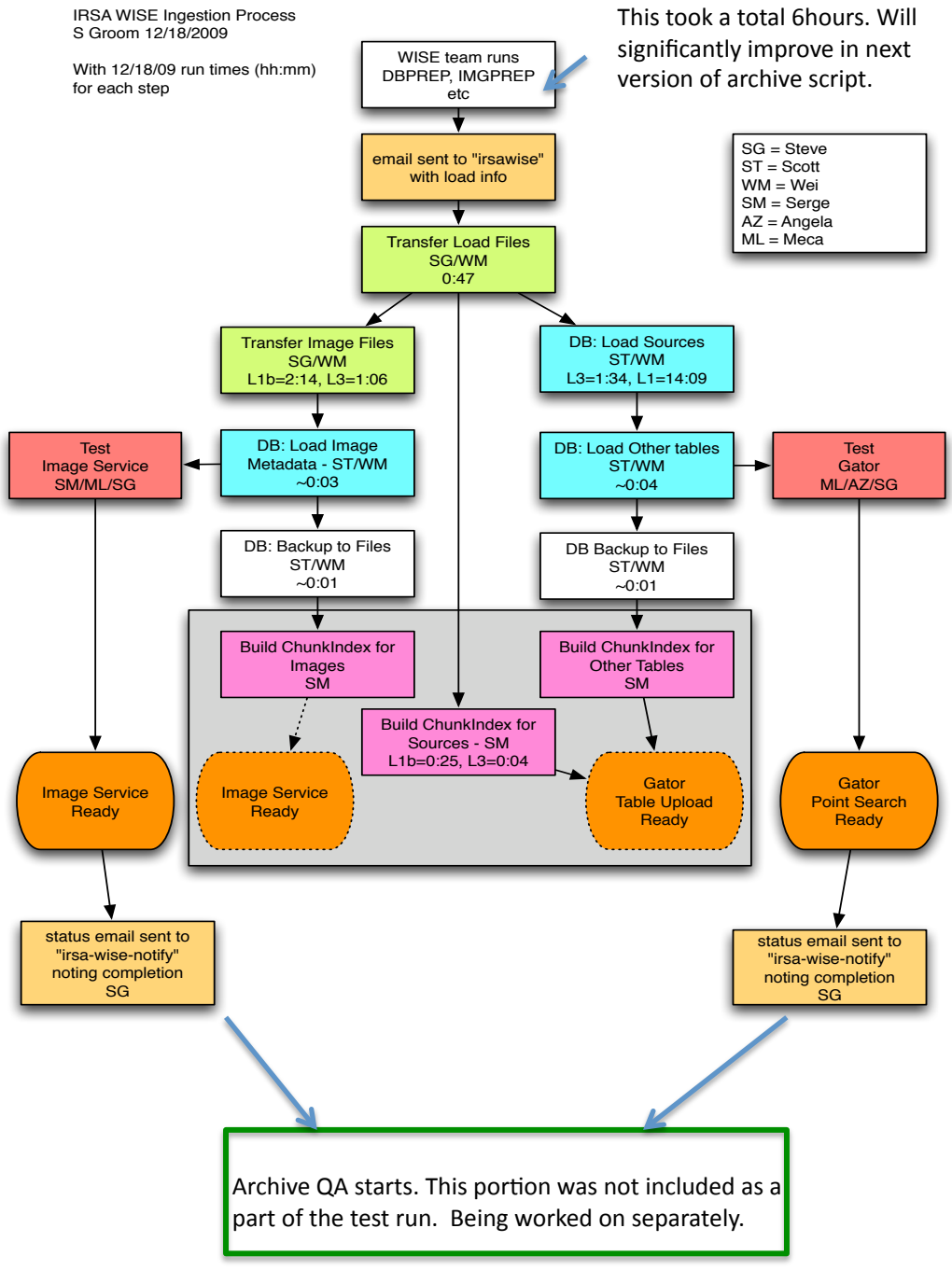
As described above, the archive loading involves the data preparation and actual loading/releasing the data. In order to ensure the WISE data being archived in a timely fashion, we plan to have the IRSA release newly loaded data on Mondays and Thursdays every week. These two dates are specified only for the data being accessible by the WISE science team. To meet these deadlines, the WISE team will need to deliver prepared datasets to the IRSA team twice a week. The tentative plan is for the WISE OPS team prepares and delivers the data to the IRSA on Mondays and Wedsndays. The dates and time for the actual data loading into the archive can be flexible and not necessarily tied to these two data releasing dates. For example, the actual archive loading could happen on Friday, then the release to the users on Monday. With this fixed schedule, we will plan the required activities prior the IRSA database loading. This will also give enough time for the WISE archive and QA team to validate the newly loaded data in the IRSA database.

3 TIMING REQUIREMENT FOR ARCHIVE LOAD

Because the archive needs to load twice a week, we want to make sure that for each loading, we can complete all tasks within a turn-around time of 3 days. Here is the summary of the time estimates for each task shown in Figure 1.

Because the QA and the pipeline processing are done continuously, the timing requirement mostly relevant to the archive is the data preparation and actual loading into the IRSA database. We have done a simulated archive end-to-end loading on December

17-18, 2009. This excise has provided the rough estimates of how much time the data preparation and loading will take. The detailed report on this end-to-end test has been documented in WSDC D-T-021. Figure 3 shows the break down of each step.



1. The time took to prepare the data (archive loading script run time) was initially 6 hours. With the newly revised code, the run time has shortened to (2-3) hours.

2. The WISE OPS will make a detailed weekly operation schedule, and the archive team will make sure that all of the required preparation work for each loading is planned in a timely fashion and can be executed on time.
3. IRSA database (GATOR) loading time: We have tested the speed of database loading. To load 66 million rows of source tables, it took only 14 hours to complete. This size table corresponds to 30 orbits of single frame data, about 2 days worth of data.
4. Fits image transfer: Without actively setting the transferring speed, the fits image transfer takes about 3 hours to transfer 266GB of fits images (level 1b and level 3 images) from the WISE computer servers to the IRSA computer disks. This data volume is roughly two days data from the normal operation. We note that the time for the IRSA data transfer/loading will be shortened after the hardware improvement is done.

5 TIME LINE FOR ARCHIVE LOADING AFTER THE WISE LAUNCH

The WISE was launched on December 12, 2009, and has gone through the In-orbital-Check (IOC) successfully. The nominal survey started on January 14th, 2010. Currently the WISE pipeline is being tuned to produce the optimal reduced data products.

We have tested the archive loading script and loaded the 10 scans of processed data into the IRSA database. During this testing, we have refined the archive loading scripts, and corrected the data dictionaries etc small problems. The WISE GATOR and the image archive have been released to the WISE science team for testing. The second load test with 50 scans is scheduled for January 22, 2010. The main purpose is to see how we deal with a large data volume, and resolve any remaining associated issues.

Between Jan. 15th and Feb. 28th, all of the processed data is regarded as provisional processed data. Once the the provisional processing pipeline parameters are finalized, we will start the archive loading. As of January 19th, the start of the loading date is not known yet. We expect we will need to do 3 loading per week initially to catch up with the accumulated data. Starting on March 1st, re-processing starts and this version of processed data is expected to ultimately replace the archive, provisional processed data. Therefore, after March 1st , we will need to have one extra loading per week in order to archive the reprocessed data taken earlier.

In addition, new improvement on the archive functionalities will be implemented after we start to make routine loading regularly.