# HAIO Performance Requirements

## Document Approval

| Approved by | Organization | Signature | Date |
|---|---|---|---|
| Leo Metcalfe | HSC | Approved by e-mail. | November 17, 2011 |
| Eva Verdugo | HSC | Approved by e-mail. | December 9, 2011 |
| Jorgo Bakker | HSC | Approved by e-mail. | November 17, 2011 |
| Pedro Osuna | SAT | Approved by e-mail. | January 17, 2012 |

# Table of Contents

# 1  Introduction

The purpose of this document is to specify performance requirements for the machine interface to the Herschel Science Archive (HSA), called the Herschel Archive InterOperability (HAIO).

The HAIO is accessed by a number of different systems: As part of data processing by the Technical Assistants at the Herschel Science Center, whether this is standard processing of each operational day as it is downloaded, or bulk reprocessing of all existing data in the HSA. The HAIO is accessed by processes performing on-demand processing as requested by external users. And the HAIO is accessed across the external internet by user installations of HIPE. A significant fraction of HAIO accesses across the internet is done by instrument control centres (ICCs) to retrieve the data from their instrument, following days on which their instrument was active. This is a process referred to as "bulk product transfer" (see [RD-3]).

All these systems evolve and the amount of data in the HSA is growing. On one hand, the various systems accessing the HAIO may at some point put an excessive load on the HAIO, and on the other hand, as usage and data content grow, the HAIO may not provide sufficient performance for the accessing systems to meet their performance requirements. This document has been prepared to keep these factors in check and to be able to know where, in which system, performance problems shall be addressed, should they arise.

# 2  Scope

This document captures only the performance requirements on the HAIO. Functional requirements and non-functional requirements other than those on performance (e.g. reliability, robustness, maintainability and security) shall be covered by the HSA URD [AD-1].

# 3  Applicable and Reference Documents

## 3.1  Applicable Documents

Applicable documents provide constraints that apply to the HAIO as well, though some of the requirements may be implemented by other HSA sub-systems, such as the Herschel User Interface.

**AD-1.** Herschel Science Archive Requirements Document, issue 1.0, 25 July 2008, Herschel-HSC-DOC-1081

**AD-2.** HCSS User Requirements Document, issue 2.4, 25 January 2008, LiveLink ID 22032

## 3.2  Reference Documents

**RD-1.** Herschel Data Access RRF Report, issue 1.0, 21 November 2007, LiveLink ID 2794367

**RD-2.** HSA-HCSS ICD, issue 1.5, 11 May 2011, LiveLink ID 2801712

**RD-3.** Herschel Bulk Product Transfer ICD, issue 1.2, 1 October 2009, HERSCHEL-HSC-ICD-1083

### 3.3 LiveLink ID

The LiveLink ID specified for some documents can be used to retrieve the document using the following URL: http://www.rssd.esa.int/llink/livelink/open/[LiveLink ID]

# 4 Terms, Definitions and Abbreviated Terms

This document uses the following terms and definitions.

### 4.1 Domain Terminology

The following terms are used to described the objects encountered in the specific domain:

- **Products and URNs**: A product is the most basic data item in the HSA. In the HSA, products are represented as FITS files (possibly with the exception of a few specific cases). Each product is uniquely identified by its URN. Every FITS file stored in the HSA is a kind of product.

- **Modeled and non-modeled metadata**: The HSA stores products. These products contain data and information describing the data, which we refer to as metadata. Some of this metadata is stored in a data-model for more efficient searching, and some of the metadata is not modeled.

- **Product descriptors**: Information about the product, such as the URN of the product. This information is not part of the product itself, in contrast to the product metadata.

- **Version and Version Track**: A product can be stored in the HSA, updated, and stored again. These products can be stored as two different versions of the same object. Products that are different versions of the same object are said to form a version track.

- **Standard Processing**: The telemetry from every operational day is processed using the data reduction pipeline, as soon as possible after it arrives at the Herschel Science Centre. Running the data reduction pipelines on this data and storing it in the HSA is called "standard processing", or also "standard product generation" (SPG).

- **Bulk Reprocessing**: After the data reduction pipeline software is updated, it can be worthwhile to reprocess all the existing data in the HSA and to store the reprocessed data as a new version of the existing data. This is called "bulk reprocessing" or also "bulk product generation" (BPG).

### 4.2 Terminology for HAIO requests

The following terminology is used to specify requests that are being sent to HAIO.

- **Product Retrieval**: The retrieval of an actual product by the HAIO from the appropriate HSA sub-system and the download to the client of HAIO that made the product retrieval request. This corresponds to MSG-2.2-001 in [RD-2] with a single URN and parameter "protocol=HTTP".

- **Multiple Product Retrieval**: Product retrieval of 50 different products using a single request. This is also referred to as "virtual tar".

- **Query**: Any request for product descriptors based on a set of criteria.

- **Standard Query**: A query based on 5 different non-modeled metadata items. The query returns the descriptors for all products matching the constraints specified by the query. This is MSG-2.1-001/002 as defined in [RD-2], with a QUERY parameter containing the constraints on the metadata items and without SHOW_PROPRIETARY parameter.

- **Descriptors Query**: A query for the descriptors of one given product, using its URN, that can be used to determine if the product exists in the HSA. This is MSG-2.1-001/002 as defined in [RD-2], with a QUERY only specifying one URN: "QUERY=(HCSS_URN==[urn value])".

- **Multiple Descriptors Query**: A descriptors query specifying URNs of 50 different products.

- **Last Version Query**: A query to determine the latest version of a product, given a version track identifier. This is MSG-2.1-007/008 defined in [RD-2] with RESOURCE_CLASS=HCSS_TRACK_ID.

- **Product Classes Query**: A query for the list of all classes of products in the HSA. This is MSG-2.1-003/004 in [RD-2]: RESOURCE_CLASS=HCSS_CLASS_TYPE with no further parameters.

## 4.3  Performance Requirements Definition Terminology

The following terms are used for the definition of the performance requirements.

- **Response Time**: The response time for a single request is the elapsed time between receipt of a request by the HAIO and the start of the download of the response. An HAIO response can contain a product, a set of products or a list of product descriptors. The time required to download the response is excluded from the "response time" of that request. When putting a requirement on response time ("the response time shall be less than..."), we refer to the 95th percentile of all response times in a given period. This means that in any given period, at least 95% of all response times shall be less than the number quoted.

- **Light Load**: The system load generated by typical usage by external users, plus standard processing. Specifically we define this as the HAIO handling less than 200,000 queries and less than 200,000 product retrievals per 24-hour period.

- **Heavy Load**: The system load generated by bulk reprocessing, standard processing and typical usage by external users, all combined. Specifically this is defined as the HAIO handling less than 1 million queries and less than 600,000 product retrievals per 24-hour period.

## 4.4  Abbreviated Terms

The following is a list of acronyms used in this document:

HIPE        Herschel Interactive Processing Environment

HAIO        Herschel Archive InterOperability

HSA         Herschel Science Archive

# 5  Software Overview

For an overview of the software and usage scenarios, please refer to the report of the Data Access RRF ("rapid response force") [RD-1]. Also the HSA-HCSS ICD will give the reader an idea of how the system is expected to be used [RD-2].

# 6  Performance Requirements

**HAIO-UR-0001**      **Product Retrieval Response Time under Light Load**

The Response Time of the HAIO for a Product Retrieval request shall be less than **150 milliseconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0002**      **Product Retrieval Response Time under Heavy Load**

The Response Time of the HAIO for a Product Retrieval request shall be less than **300 milliseconds**, when the HAIO is experiencing Heavy Load.

**HAIO-UR-0011**      **Multiple Product Retrieval Response Time under Light Load**

The Response Time of the HAIO for a Multiple Product Retrieval request shall be less than **2 seconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0012**      **Multiple Product Retrieval Response Time under Heavy Load**

The Response Time of the HAIO for a Multiple Product Retrieval request shall be less than **4 seconds**, when the HAIO is experiencing Heavy Load.

**HAIO-UR-0101**      **Standard Query Response Time under Light Load**

The Response Time of the HAIO for a Standard Query shall be less than **1 seconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0102**      **Standard Query Response Time under Heavy Load**

The Response Time of the HAIO for a Standard Query shall be less than **5 seconds**, when the HAIO is experiencing Heavy Load.

**HAIO-UR-0111**        **Descriptors Query Response Time under Light Load**

The Response Time of the HAIO for a Descriptors Query shall be less than **150 milliseconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0112**        **Descriptors Query Response Time under Heavy Load**

The Response Time of the HAIO for a Descriptors Query shall be less than **300 milliseconds**, when the HAIO is experiencing Heavy Load.

**HAIO-UR-0121**        **Multiple Descriptors Query Response Time under Light Load**

The Response Time of the HAIO for a Multiple Descriptors Query shall be less than **2 seconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0122**        **Multiple Descriptors Query Response Time under Heavy Load**

The Response Time of the HAIO for a Multiple Descriptors Query shall be less than **4 seconds**, when the HAIO is experiencing Heavy Load.

**HAIO-UR-0131**        **Last Version Query Response Time under Light Load**

The Response Time of the HAIO for a Last Version Query shall be less than **200 milliseconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0132**        **Last Version Query Response Time under Heavy Load**

The Response Time of the HAIO for a Last Version Query shall be less than **500 milliseconds**, when the HAIO is experiencing Heavy Load.

**HAIO-UR-0141**        **Product Classes Query Response Time under Light Load**

The Response Time of the HAIO for a Product Classes Query shall be less than **150 milliseconds**, when the HAIO is experiencing Light Load.

**HAIO-UR-0141**     **Product Classes Query Response Time under Heavy Load**

> The Response Time of the HAIO for a Product Classes Query shall be less than **150 milliseconds** (same as under light load), when the HAIO is experiencing Heavy Load.

We note that we do not specify any performance requirements for loads higher than the heavy-load scenario. Requirements on reliability will be sufficient.

## 6.1 Summary

The following table summarizes the requirements.

| ID | Operation | Light Load | Heavy Load |
|---|---|---|---|
| 0001/0002 | Product Retrieval | 150 ms | 300 ms |
| 0011/0012 | Multiple Product Retrieval | 2 s | 4 s |
| 0101/0102 | Standard Query | 1 s | 5 s |
| 0111/0112 | Descriptors Query | 150 ms | 300 ms |
| 0121/0122 | Multiple Descriptors Query | 2 s | 4 s |
| 0131/0132 | Last Version Query | 200 ms | 500 ms |
| 0141/0142 | Product Classes Query | 150 ms | 150 ms |

# 7 Validation

This section describes how the system can be validated against each of the requirements from section 6.

## 7.1 Requirements on Product Retrieval

We begin by constructing a list of URNs of small products that exist in the HSA. This list shall be sufficiently large such that repetition of URNs is not common. The products of choice will be ObservationContext products, which have typical size of less than 100 kB. This allows network transfers to be insignificant in our measurements.

If we select a URN from the list and retrieve it from the HSA using the Unix command "wget", we can measure the (real) elapsed time using the Unix command "time". We do this for a 24-hour period every 5 minutes, giving us 288 measurements for the day. If less than 15 of these measurements (5%) exceed the time set by the requirement, the requirement is met.

An alternative is to measure the response times on the server, but it would have to be the "outermost" server, i.e. the server that the clients connect to directly.

For the requirements on Multiple Product Retrieval, one simply bundles the required number of URNs in a single request.

## 7.2  Requirements on Queries

Along the same lines as validation of product retrieval requirements, we construct queries and measure their responses time using the Unix commands "wget" and "time". To get a reliable number, we must do this significantly more often than 100 times in a given period. From the recorded measurements we determine the 95$^{th}$ percentile and compare to the requirement.

The standard query can ask for an observation in a fixed OD range, for a given instrument and given "creator" (a valid value for the creator metadata item is "SPG v1.0").

The Descriptors Query will retrieve the descriptors for a URN with a random element – these should be a mixture of URNs that for which a product exists and URNs for which no product exists in the HSA.

The last version query constructs a version track ID from a list of version track IDs that exist in the HSA and requests the last version of the selected track ID.

The product classes query simply retrieves the list of product classes in the HSA.

## 7.3  Requirements on Descriptor Queries

For the requirements on descriptor queries, one can apply the same strategy, only generating random URNs using a standard prefix. For example: "urn:hsa:herschel.ia.obs.ObservationContext:" followed by a random number between 0 and 10,000. It is important to check a given URN only once as the second time the response will be faster due to caching if a request is repeated soon after the first request (this is an observed effect).

# 8  Traceability

The requirements on query performance implement the following requirements from the HSA URD [AD-1]:

- **HSA-UR-0480**: The system shall return the results of a query on product metadata on 10 parameters in less than 5 seconds (one second is desirable).

- **HSA-UR-0490**: For Trend Analysis, it shall be possible to perform a query on a certain product type and return the data values for up to 10 parameters from a time period of up to 1 week in less than 30 minutes (this is a goal, since the query duration depends on the data volume queried).

All requirements in this document combined implement the requirement from the HCSS URD [AD-2]:

- **HCSS-UR-3.2-0180** [AD-2]: In average, the HCSS shall process the TM associated with 12 Operational Days of 24 hours (nominal) each, and produce the corresponding end-products (scientific, auxiliary data and quality data) in 24 hours.

  The goal should be to process 24 hours of TM data in 15 minutes.