The RHESSI Experimental Data Center

Pascal Saint-Hilaire[1,2], Christoph von Praun[4], Etzard Stolte[3], Gustavo Alonso[3], Arnold O. Benz[1] and Thomas Gross[4]
[1] *Institute of Astronomy, ETH Zürich, CH-8092 Zurich, Switzerland* (shilaire@astro.phys.ethz.ch)
[2] *Paul Scherrer Institute, CH-5232 Villigen, Switzerland*
[3] *Institute of Information Systems, ETH Zürich, CH-8092 Zurich, Switzerland*
[4] *Laboratory for Software Technology, ETH Zürich, CH-8092 Zurich, Switzerland*

Abstract.
The RHESSI Experimental Data Center (HEDC) at ETH Zurich aims to facilitate the use of RHESSI data. It explores new ways to speed up browsing and selecting events such as solar flares. HEDC provides pre-processed data for on-line use and allows basic data processing remotely over the Internet. In this article, we describe the functionality and contents of HEDC, as well as first experiences by users. HEDC can be accessed at http://www.hedc.ethz.ch. Additional graphical material and color pictures are available on the CD-ROM accompanying this volume.

1. Introduction

The Reuven Ramaty High Energy Solar Spectroscopic Imager (RHESSI) images the Sun in the 3-17000 keV range (soft and hard x-rays, as well as gamma rays) with unprecedented spatial, temporal and spectral resolutions. RHESSI's main objective is to deepen our understanding of energy release and particle acceleration taking place in solar flares. Furthermore, it provides spectroscopic information on cosmic gamma-ray bursts, and even image such extra-solar objects as the Crab Nebula (Lin et al., 2002).

RHESSI's nine rotating modulation collimators (RMC) modulate the incoming X-rays on the detectors at the end of each RMC. As each RMC's grid pair has different pitches, a set of nine Fourier-like components are obtained, and are used to reconstruct the original image (Hurford et al., 2002). As each incoming photon is tagged in time, energy and RMC, these photons can be binned in a specified energy band for image reconstructions, lightcurves, and spectra.

A good proportion of existing astrophysical databases concentrate on archiving raw data from one or a few instruments such as Yohkoh, TRACE, RHESSI, or ground-based observatories such as Phoenix-2. Examples are the Solar UK Research Facility (SURF[1]) and the TRACE Data Center[2]. Other databases offer a selection of data (e.g., synoptic maps) from several observatories and provide usually easier or faster access to these data than the primary archives. Examples are the Base de données Solaire Sol 2000 (BASS2000[3]) (Mendiboure, 1998) and the SOHO Summary[4] and SOHO Synoptic[5] search engines.

Most other observatories' data are stored in an immediately usable form, albeit not (fully) calibrated, but suitable for quick perusal. RHESSI data poses the problem that it must be reconstructed to be of any use, much like Hard X-ray Telescope (HXT) data, but on a larger scale.

Image reconstruction can take from less than a minute (for a basic *back projection*) to a several hours (for a tedious *pixon* reconstruction) on a state-of-the-art workstation. Image reconstruction must be done by each data analyst and requires significant hardware and software resources, including the commercial product Interactive Data Language (IDL).

[1] http://surfwww.mssl.ucl.ac.uk/surf/
[2] http://vestige.lmsal.com/TRACE/DataCenter
[3] http://bass2000.bagn.obs-mip.fr/New2001/Pages/page_acceuil.php3
[4] http://sohowww.nascom.nasa.gov/cgi-bin/summary_query_form
[5] http://sohowww.nascom.nasa.gov/cgi-bin/synop_query_form

Each flare may require many images, at different times, energy intervals, accumulation time intervals, image sizes and resolution, etc. If we consider the importance of spectra, lightcurves and other ancillary data, we realize that large amounts of computing time are needed to create the derived (i.e. Level-1) data an observer will sift through, to home in on datasets of interest.

These preparatory, but necessary, computational activities partially overlap among each observer and project. This realization provides the starting point for the RHESSI Experimental Data Center (HEDC). Its purpose is to automatically generate an exhaustive amount of 'quick-look' data products and assemble them in an on-line data warehouse that will allow fast browsing and other services.

The 'E' in the HEDC acronym stands for 'Experimental': HEDC introduces several innovations but also serves as a platform for deepening our understanding of scientific data warehouses. HEDC is not simply another data repository, but also a database of scientifically useful derived data. Furthermore, every derived data item is accessible on-line, e.g., through the use of any Web browser.

On-line data processing (on the HEDC's servers) by users supplements the available data products. Users can add their own data products to the database, and the derived data on events of special interest thus increases in a self-organizing way. User participation will increase the scientific return of HEDC, and ultimately of RHESSI.

HEDC is a joint project of several groups at ETH Zürich: the Institute of Astronomy, the Institute of Information Systems, and the Laboratory for Software Technology (the last two are in the Department of Computer Science).

Section 2 describes the HEDC system. Section 3 explains how to use HEDC. Section 4 summarizes first experiences by users and concludes this paper. Appendix A describes the contents of the HEDC Extended Catalog (events and data products). Appendix B lists of all user-relevant query attributes on HEDC.

2. Description of HEDC

HEDC classifies its data as *events* and *data products.* An event (sometimes more specifically referred to as an *HEDC event*) can be a solar flare, a gamma-ray burst, or a terrestrial electron precipitation. The term *other event* is reserved for future extensions or events that are not yet determined. An event consists of a list of attributes (such as start and end times, total counts, peak count rates in certain energy bands, etc.) and a list of associated data products. Data product is the

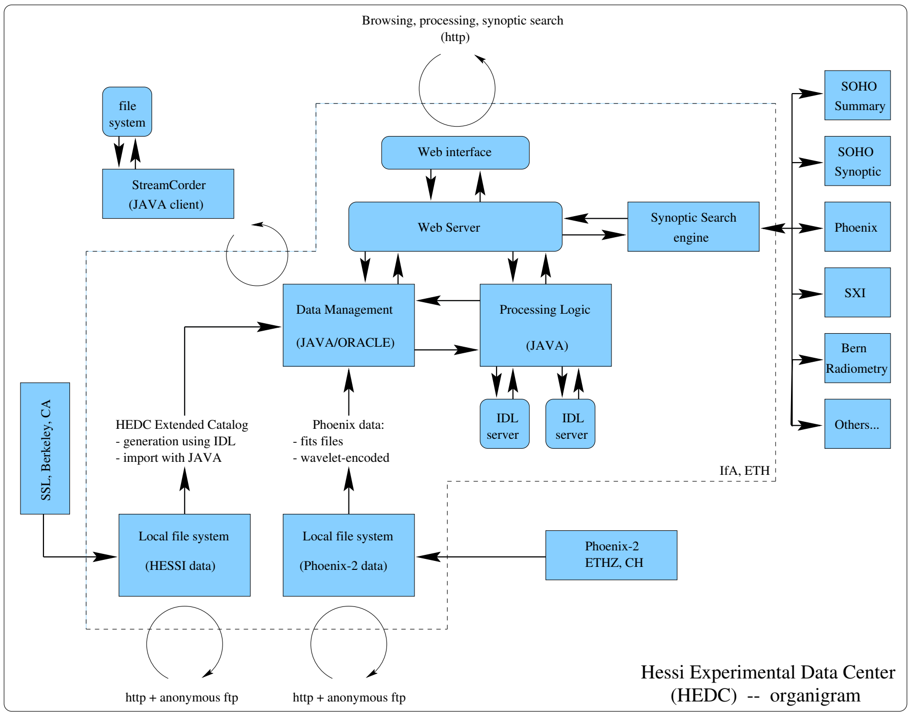

Figure 1. The HESSI Experimental Data Center at ETH Zürich

generic term that refers to all derived data: images, lightcurves, spectra, spectrograms, etc. A data product consists of a list of attributes (in this case, the more important parameters that were used to generate the derived data, such as accumulation times, energy bands, etc.) and a picture (PNG or JPEG format). The list of attributes for both events and data products are tables for database querying.

The main services provided by HEDC are:

- On-line database for events and data products;
- On-line RHESSI data processing;
- On-line RHESSI Level-0 data repository that contains the raw data;
- Other on-line services such as a Synoptic Search engine that quickly retrieves other solar quicklook data of relevance.

Figure 1 shows an organigram of the different parts of HEDC. Its principal components are:

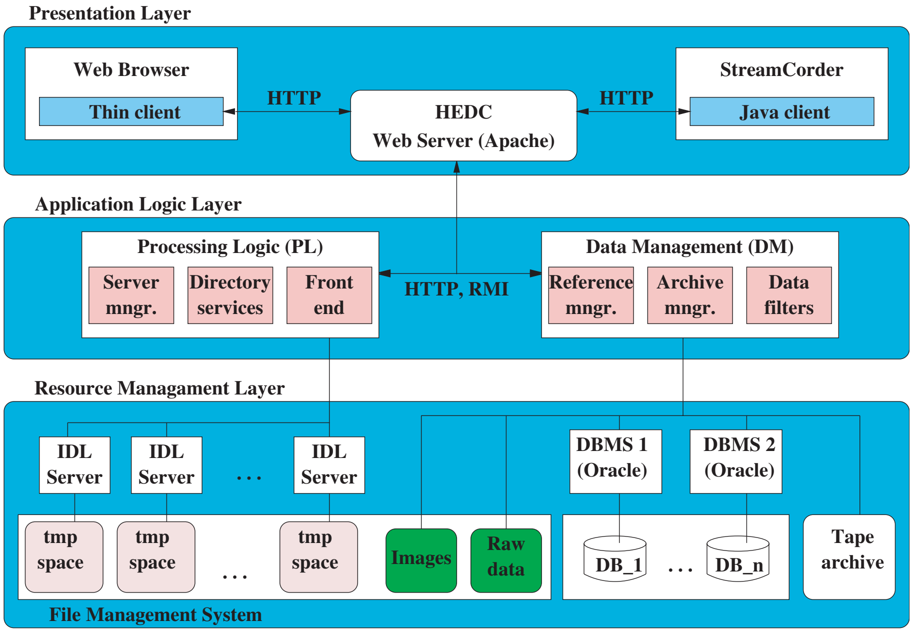

Figure 2. Architecture of HEDC

- A file system for all Level-0 RHESSI and Phoenix-2 (Messmer et al., 1999) calibrated data files;
- The HEDC Extended Catalog generation (using IDL) subunit;
- The Data Management (DM) subunit;
- The Processing Logic (PL) subunit and the IDL servers;
- The Synoptic Search engine;
- A Web server providing the main user interface;
- The StreamCorder, an alternative to the Web (browser) interface that provides more flexibility.

2.1. Architecture

HEDC is implemented as a 3-layer architecture (Figure 2), where the intermediate application logic layer relays requests for data and/or processing by *clients* (programs that run on local workstations) to an Oracle Relational Database Management System (DBMS) and a number of processing *servers* (remote computers at ETH) at the resource management layer. The application logic layer consists of two components: (1) The Data Management (DM) component takes care of all data storage issues. (2) The Processing Logic (PL) component acts as an intermediary between the DBMS and the IDL servers. Both are implemented in Java and run as stand-alone programs. HEDC can be accessed through either a Web-based client (e.g., a conventional browser)[6] or a Java-based client, the StreamCorder. With any of the two interfaces, users can query and download raw data, view data products and perform new processing steps. The Java-based client offers some functionality not available in the Web interface, e.g., tools for data visualization and system administration.

HEDC currently runs on a SUN Enterprise Server with 2 GBytes RAM, two 450 MHz processors, two hard-disk drives of 36 GBytes each, and two RAIDs of 654 and 1795 GBytes. Critical data, such as the database configuration information, is stored on RAID with tape backup, as are the derived data and the raw data files. The IDL processing servers execute on the SUN server and on Linux PCs.

2.2. On-line RHESSI data repository

RHESSI raw data (the Level-0 data) are stored in the form of FITS files, from which all RHESSI derived data are produced, using the RHESSI data analysis software[7] (Schwartz et al., 2002).

The raw science data are mirrored daily from the Space Science Laboratory in Berkeley (CA, USA) in their entirety. In keeping with RHESSI's open data policy, all RHESSI data files are publicly available via anonymous FTP[8] or by HTTP access from the HEDC home page. The raw data come in at a rate of about 1.8 GBytes per day, taking about three hours to download. The data are usually available one or two days after observation, slightly more in case of a big flare (because of the limited downlink time between spacecraft and ground stations).

2.3. HEDC Extended Catalog generation

Along with the Level-0 data files usually comes 'quicklook' derived data (included are images and a flare list, amongst other items, see the online software documentation for further details). HEDC includes those quicklook data products (called here the *standard* catalog) and adds a large amount of other derived data, giving rise to the HEDC *extended* catalog.

The raw data are scanned for events of interest: this is presently done mostly via the flare list incorporated with the quicklook data.

[6] http://www.hedc.ethz.ch/
[7] http://www.RHESSI.ethz.ch/software/
[8] ftp://hercules.ethz.ch/pub/hessi/data/

Parameters of interest for each event (total counts, peak count rates, etc.) are then extracted and stored in the DBMS as attributes for database queries. A set of data products is then generated for each such event: spectra at different times, images at different times using different energy bands, lightcurves, etc. Whenever possible, additional data relevant to the event are added, e.g., Phoenix-2 radio spectrograms, or the quicklook images or spectra.

Appendix A gives the current list of the data products computed automatically for each event. All new events and data products are inserted in the HEDC database, and may be recovered by anyone.

The whole process of creating the extended catalog, using IDL and the *SolarSoft*[9] (SSW) libraries, and inserting its elements into the database is fully automated, producing a standarized set of data products for each event.

The automatic generation of the extended catalog is still complicated by the changing SSW environment and the (as yet) lack of reliable flare positions. It will be constantly improved and completed over the next months, and derived data generated for the extended catalog will be reprocessed regularly until a final satisfactory stage is reached.

2.4. The data management (dm) component

The Data Management (DM) component handles data requests by external clients and by the processing logic (PL). It offers HTTP and Remote Method Invocation (RMI) interfaces that hide the complexity of the lowest layer and provide an abstraction from the database schema (i.e. the list of database attributes).

Web clients access the DM through the HEDC Web server. Requests are first analyzed to determine the sub-systems needed to create the response. Then the relevant data and data references are retrieved from the database and data repositories, and a dynamic HTML response page is generated. StreamCorder clients access the DM directly through RMI, so that no HTML pages need to be created.

In astrophysics, system architectures are typically kept simple to cope with the high amount of data provided. Often, access systems are based on FTP combined with simple query scripts. Within the DM component, HEDC uses a commercial object-relational DBMS to manage the meta-data describing the derived data created by users and automated routines. Using a database simplifies the design of the data center as complex tasks can be left to the database rather than implemented anew. Furthermore, it offers greater flexibility for adding new features as need dictates.

[9] http://www.lmsal/solarsoft/

Advantages of using a DBMS rather than a file system include consistent data updates with concurrent users, automatic as well as dynamic creation and maintenance of indexes, flexible query capabilities, efficient in-memory data caching for faster access and query processing, view materialization to avoid repetitive work in answering complex queries, and a flexible framework as the number of users increases and HEDC widens its scope. These are all important features that help HEDC to provide much more sophisticated capabilities than file based systems at a lower development and maintenance cost. The database also takes care of efficient disk utilization and reorganization in an automatic manner and without requiring manual intervention. This is an important feature given the amount of data involved and the high update frequency. These and many other advantages have also been observed by other projects that have followed a similar approach (see, for instance, the work related to the Sloan Digital Sky Survey (Szalay et al., 2000; Szalay et al., 2002)).

Moreover, the fact that HEDC is built upon a database has allowed us to extend the architecture in several interesting directions that would have been cumbersome to pursue if HEDC would have been based on a file system. For instance, one of the HEDC interfaces, the StreamCorder, is meant to reside in the user's computer. Under the covers, the StreamCorder contains another object-relational DBMS that mirrors the database schema at the server. As the users work and process HEDC data using the StreamCorder, raw data and derived data are cached locally, thereby greatly speeding up overall processing. Users of the StreamCorder can work disconnected from the HEDC server using the data stored locally. They can also update this local data with new data products which will be automatically uploaded into the HEDC server once the user is on-line again. The StreamCorder has been designed so that users can gradually create their own HEDC data center provided they have the storage space. Everytime a user performs an operation with the StreamCorder (executes a processing, formulates a query), the StreamCorder checks whether the data is available locally. If that is not the case, it retrieves the data from the HEDC server and caches it locally. Over time and without any effort on their part, users will find that most of the data relevant for their work are stored locally in their small version of the HEDC server. This mechanism helps both HEDC, since it reduces the load at the server, and the final users, who get a much faster access to the data relevant to them.

2.5. The processing logic (PL) component

The Processing Logic (PL) allows each user to compute data products beyond those that are part of the existing extended catalog provided by HEDC. Currently, three types of processing are supported, corresponding to the three main objects of the RHESSI software: images, lightcurves and spectra. Each processing activity can be configured with a set of basic parameters.

Users access the PL through the Web-based interface. Each user interacts with the system in a personalized session. Processing steps are specified as tasks that are handled by the system in accordance with the availability of computing resources. Processing is done in the background so that users can submit several tasks at once. Each task is an individual batch operation. This design choice has been taken to avoid user-specific resource reservations on the server. This leaves the scheduling of individual tasks unconstrained, leading to improved resource utilization.

Online processing is integrated with the DM in the following two ways: (1) At the user interface, the PL can be easily accessed while browsing so that the standard attributes of a data product from the extended catalog are copied into the *processing submission form* of the PL. (2) Processing results (PNG and FITS format) can be submitted to the DM and stored permanently into the database.

The implementation of the PL is based on an object-oriented software framework. The system has been designed to easily accommodate changes and additions in the supported types of processing activities; the task control and scheduling system is strictly separated from application-specific issues, such that more than 95% of the code is independent of the three currently implemented processing activity types. The structure of the system is based on service modules that execute independently and allow distributed task execution in a network of workstations. The session and task management overhead is marginal compared to the cost of data transfers and computation.

Duration of individual tasks can vary significantly. We limited the total CPU time per task to 20 minutes and constrained the set of admissible input parameters (i.e., no *pixon* image reconstructions). Those limitations may be alleviated in the future, depending on usage of available computing resources.

2.6. Other services

2.6.1. *Synoptic Search*

The synoptic search subsystem serves to quickly browse through ancillary data related to a particular event in remote astro-archives. The data obtained are usually daily GIF or JPEG images.

The query mechanism resembles a Web-crawler: first, online requests are issued to several remote archives in parallel; then, the results are collected, grouped and displayed to the user.

This service operates largely independent from other subsystems of HEDC. The service is best effort (if a query to a remote archive times out, no results are available). This light-weight approach of rendering synoptic data accessible through HEDC has proved to be practical and robust. In its current configuration, six remote archive sites are searched, including the SOHO summary/synoptic data archive and the RAPP archive at ETH Zurich. Due to its flexible software architecture, additional Internet-archives can be easily integrated.

Data in FITS format are better suited for data analysis, but are not displayed by the usual Web browsers, hence slowing down the search for datasets of interest. The RHESSI Synoptic Data Archive[10] (Zarro et al., 2002) provides an ample amount of such FITS data from other observatories, concentrated around RHESSI flare times.

2.6.2. *The StreamCorder*

HEDC is not only accessible through a Web browser, but also through the StreamCorder (see Figure 3), which must be installed locally[11]. Except for some performance-sensitive hardware-dependent routines, it has been completely implemented in Java. The architecture is extensible and modules are loaded according to the current data context. Modules may access core services such as stream management, request queues and local analysis programs. Currently available modules support browsing and download of all data types stored in HEDC, allow local and remote processing and offer administrative tools. During RHESSI data processing, the StreamCorder coordinates the asynchronous download, caching, decoding and processing of the data. A local database transparently caches query results and manages downloaded files. The local DBMS schema and the structure of the local file-system archive are identical to the ones on the server. Thus, offline work is possible.

[10] http://orpheus.nascom.nasa.gov/∼zarro/synop/
[11] http://www.hedc.ethz.ch/release/

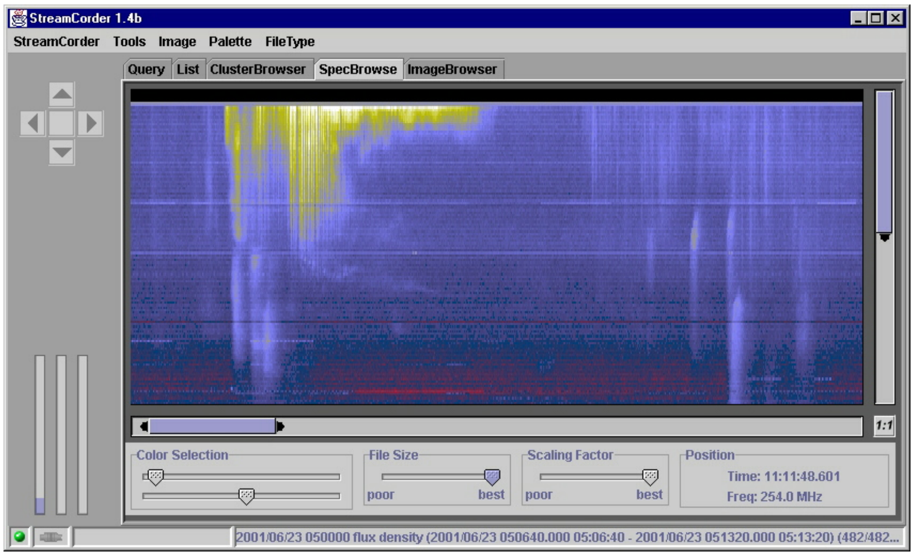

Figure 3. The StreamCorder is an alternative to the Web interface. It can search and display all HEDC data products. Here, it is displaying a wavelet-encoded radio spectrogram from Phoenix-2.

2.6.3. *Phoenix-2 archive*

HEDC also holds the Phoenix-2 radio spectrometer archive, both in FITS format and in a wavelet-encoded format, the latter for speedy spectrogram viewing with the StreamCorder.

3. Using HEDC

The term *browsing* is used in this paper to refer to one of these activities: making database queries (either by event or by data products), exploring the result set by going through the links, making another query (perhaps a finer one, or an entirely new one), etc.

The standard workflow model shown in Figure 4 for users of HEDC is to browse back and forth for events and/or data products, eventually to make new data products online, and to add them to the HEDC database. Once a user has zeroed in on a dataset of interest, he or she will have to make a thorough scientific analysis on his or her workstation, perhaps downloading some of the images previously made on HEDC. Of course, a user may decide to use only parts of HEDC for browsing, processing, or synoptic searching.

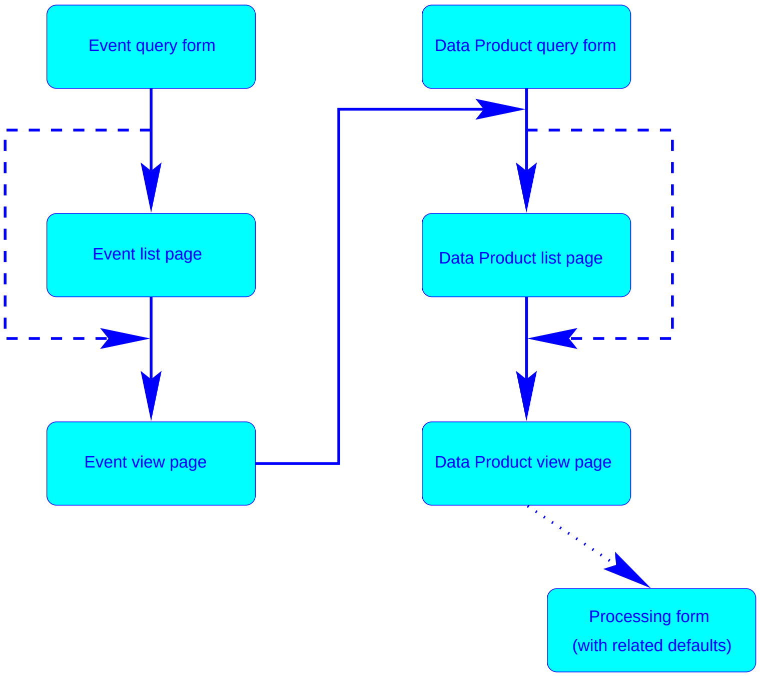

Figure 4. Browsing the HEDC using the Web interface. There are two possible entry points: the event query form or the data product query form. After submitting the appropriate query form, a list page containing the result set appears. Clicking on one of its elements leads to the view page, where the full set of database attributes can be examined, as well as a picture in the case of data products. If a query results in a single set, then the list page is bypassed. To the event view page is appended a data product list page, containing the data products associated with that event. The data product view page also has icons to access the processing form: using those instead of directly going to the processing form via the main link on the left of the Web page has the advantage that most of the processing form's attributes are already defaulted to those of the data product just examined.

3.1. Browsing with the Web interface

The HEDC Web interface currently offers two types of querying: either by event or by data products (Figure 4). Both possibilities can be done using either a standard Web form, or a more advanced one.

The *standard* Web form is intended for use by casual users. It is simple and intuitive: there should be no need to consult the on-line documentation. However, only a handful of query fields are available. The *advanced* form should be used by people more familiar with the system, and offers the full range of user-accessible database queries.

Once the query has been submitted, a *list page* appears. There is a current limit of 100 entries on this list. Each entry represents a different event or data product, with a few attributes (time intervals, energy bands, imaging algorithm, etc.) to guide the next choice of the user. Each entry is actually a link to an event (or data product).

Clicking on an event will lead the user to an *event view page*, an HTML page that displays all the event's database attributes (e.g., count rates), as well as a list of all associated data products. A data product view form also lists that data product's attributes (imaging algorithm, etc.), as well as a picture (i.e. 2-D plots or 3-D intensity maps).

A comprehensive set of examples are available in the online documentation.

All RHESSI images on HEDC may be downloaded in FITS format (as produced by the RHESSI software *fitswrite* method), by clicking on the appropriate icon in the *data product view page*.

3.2. Processing with the Web interface

Whereas browsing is open to the public, an account is needed to perform processings on HEDC (account requests can be made online).

Once a filled processing form is submitted, a *job list* appears. It is a listing of all job requests that were sent with their states (pending, running, finished, failed). Submitted jobs do not share CPU time: rather, each job is queued or executed as fast as possible using one of the available CPUs.

If a job successfully ends, an icon that displays the resulting picture appears in the job list. Clicking on this icon allows the user to view the full picture, as well as to obtain relevant database attributes and other items pertaining to it. One such item is the IDL output, particularly useful in understanding the cause of a failed job. The current setup allows a maximum of 10 jobs at any time per user.

Jobs stay on the job list until the user logs out, up to a maximum of one week. Clicking on the 'update' button will update the job list to its latest status.

3.2.1. *User events: folders for users' data products*

Once a new data product has been generated and is being viewed by a user, it is possible to store it permanently in the HEDC database. Each user-made data product must be saved in a *user event*, which is just another event on HEDC, and serves as folder (or directory) for users' data products. In this manner, individual users can create several different folders, one for each of their projects, and put in them whatever data products they process on HEDC. User events do not have any attributes, except for a *code* starting with the username. This means that a query for *events* using time intervals does not reveal user-made events, even if the user-made data products stored inside are within that time interval.

If one is interested in all data products ever generated for a given time interval, one should browse using the *data product query form*. Both user-made and HEDC-made data products are shown (HEDC-made data products have a code similar to their parent event, always starting with 'HX').

3.3. Synoptic engine

Using the HEDC's Synoptic Engine is straightforward: a user enters an approximative date and time of interest and submits the request. A list of available links to pictures appears. Choosing a longer time-out than the default value may result in more links found.

3.4. The StreamCorder

The StreamCorder provides the user with most of the previously described functionalities but in a more flexible manner than the browser-based interface. As it is a client side application (it runs on a user's local workstation), it uses local resources for processing, in contrast to the browser-based interface that employs the HEDC server. This offloads work from the server and allows for faster interactions with the system in case of repetitive queries and processings.

The StreamCorder also implements additional features that would be too expensive (in terms of CPU requirements) to provide in a centralized manner. Some examples are (1) a 'MovieCordlet' extension allows users to rapidly view a sequence of images made on HEDC, (2) the 'Spectrum Browser' enables users to look at wavelet-encoded Phoenix-2 radio spectrograms, allowing for fast exploration of the raw data (Stolte et al., 2002), (3) IDL sessions can be run locally or remotely, (4) a 'Cluster Browser' allows users to visualize the density of *event* or *data product* population in a phase space, where a 'phase' corresponds to any numerical database attribute (Stolte et al., 2002). The StreamCorder is fully operational, although slight improvements are still being applied to increase ease-of-use.

4. First experiences and conclusions

From a user's standpoint, HEDC addresses the major software constraints that create a barrier to starting the analysis of RHESSI data: (1) the purchase, maintenance, and installation of IDL, (2) the installation, configuration, regular update of software unique to RHESSI,

(3) the need to learn the detailed use of RHESSI software. Users can easily and quickly examine a huge variety of RHESSI data products, or create their own, with only a Web browser. As a side effect, HEDC is appreciated by those working at home, and who do not have available the necessary software or who lack sufficient transmission bandwidth to download raw data files.

As HEDC uses the SSW/RHESSI software to produce data products, the results are exactly the same as if they were produced from a standard IDL session. Relying on the SSW/RHESSI software makes it easy to compare results obtained at HEDC with other results. Furthermore, the substantial effort in writing (and evolving) this software is not duplicated. Of course, HEDC is therefore tightly coupled to the overall SSW/RHESSI software development and exposes the user to the perils of a dynamic software environment. However, the benefits obtained now (and in the future, after the final release) outweight any temporary glitches.

As of the beginning of August 2002, HEDC contains more than 10000 data products of over 400 flares. Having a database to classify all pertinent RHESSI data products instead of a file system-based archive is a big advantage: a single event can warrant so many data products that users get lost trying to sort them again if they rely on a standard file system. HEDC allows for quick, easy searches.

For the astrophysics researcher, the determination of flare position has been found to be most useful. Many small flares not in the catalog are located by users and stored. The demand for this function will greatly increase once the aspect solution will be more reliable. Also in high demand are mission-long lightcurves and the visualization of the data in the observing summaries of satellite orbits.

Acknowledgements

We thank the RHESSI software team for continuous encouragements and support, in particular Brian Dennis, André Csillaghy, Jim McTiernan, Richard Schwartz, and Kim Tolbert for their help, explanations, feedback, and goodwill.

The RHESSI work at ETH Zurich is supported, in part, by the Swiss National Science Foundation (grant nr. 20-67995.02) and ETH Zürich (grant TH-W1/99-2).

References

Handy B. *et al.*, 1999: *Solar Phys.*, **187(2)**,229
Hurford, G. et al., 2002: *Solar Phys.*, this volume
Lin R.P. et al., 2002: *Solar Phys.*, this volume
McTiernan J.M. et al., 2002: *Solar Phys.*, this volume
Mendiboure C., 1998: Second Advances in Solar Physics Euroconference: ASP Conf. Series **155**, 302
Messmer P., Benz A.O., Monstein C., 1999: *Solar Phys.*, **187 (2)**, 335
Schwartz R.A. et al., 2002: *Solar Phys.*, this volume
Stolte E., Alonso G., Optimizing Scientific Databases for Client-Side Proccessing. proceedings of the VIII Conference on Extending Database Technology (EDBT), Prague, Czech Republic, March 2002.
Stolte E., Alonso G., Efficient Exploration of Large Scientific Databases. Proceedings of the 28th International Conference on Very Large DataBases (VLDB), Hong Kong, China, August 2002.
Szalay A. S. , Gray J., Thakar A., Kunszt P.Z., Malik T., Raddick J., Stoughton C., van den Berg J., 2002: The SDSS SkyServer - Public Access to the Sloan Digital Sky Server Data. ACM International Conference on Management of Data, SIGMOD 2002.
Szalay A. S., Kunszt P.Z., Thakar A., Gray J., Slutz D.R., 2000: Designing and Mining Multi-Terabyte Astronomy Archives: The Sloan Digital Sky Survey. ACM International Conference on Management of Data, SIGMOD 2000.
Zarro, D. M.; Tolbert, K.; Dennis, B. R., 2002: *Solar Phys.*, this volume

Appendix

A. HEDC Extended Catalog contents

This appendix describes the current state of what is generated and stored on HEDC. It is liable to change. Consult the on-line documentation for the latest updates. Currently, the generation of the HEDC extended catalog is done about a week after observation by RHESSI. Later reprocessings will occur periodically and incrementally, following improvements or additions to the catalog generator, or major modifications to the raw data or the flare list. The newest, reprocessed versions of HEDC events and their associated data products will replace previous versions. Of course, user-made events and data products will never be reprocessed.

A.1. Detection of events

Currently, only solar flares and some 'other' flares (i.e. with parameters still undefined) are being looked for and generated. Later, this might be extended to gamma-ray bursts and electron events.

Solar flares are given by the flare list attached to the Level-0 data. Basically an increase in photon count rates in the 12-25 keV energy band is looked for. The signal must also be strongly modulated in RHESSI's two coarsest detectors (number 8 and 9). See the RHESSI Data Analysis Software pages[12] for more details on this.

'Other flares' are those enhancements in the count rates as seen by HEDC or the flare list, and for which no other classification was (yet) found.

After an event is detected, and its type determined, a set of attributes is determined which characterize the event for later database queries.

A.2. Determination of event attributes

Attributes for each events (such as start and end times, total counts, peak count rates in certain energy bands, SAA and eclipse flags, etc.) are determined as each event is generated. Those attributes can be used as search fields during database queries.

Appendix B gives a full listing of HEDC event attributes.

A.3. Data Products automatically generated with each event

For all events:

- lightcurves of the whole event, in different energy bands.
- three spectra in the 3-2500 keV range, with one minute accumulation time. One done at peak time, one midway between start time and peak time, and one midway between peak and end time.
- images made from Observing Summary (McTiernan et al., 2002) data: count rates in different energy bands; RHESSI trajectory on a Mercator projection of Earth; modulation variance lightcurves; flags; geomagnetic latitude (Figure 5 and 6).
- RHESSI spectrograms are generated. If possible, they are also superimposed with radio spectrograms from Phoenix-2 (Figure 7). Both are backgound-subtracted.
- background-subtracted time series of the 25-50 keV over 12-25 keV counts ratio for the whole event.

Additionally, for 'solar flare' events only:

[12] http://www.RHESSI.ethz.ch/software/

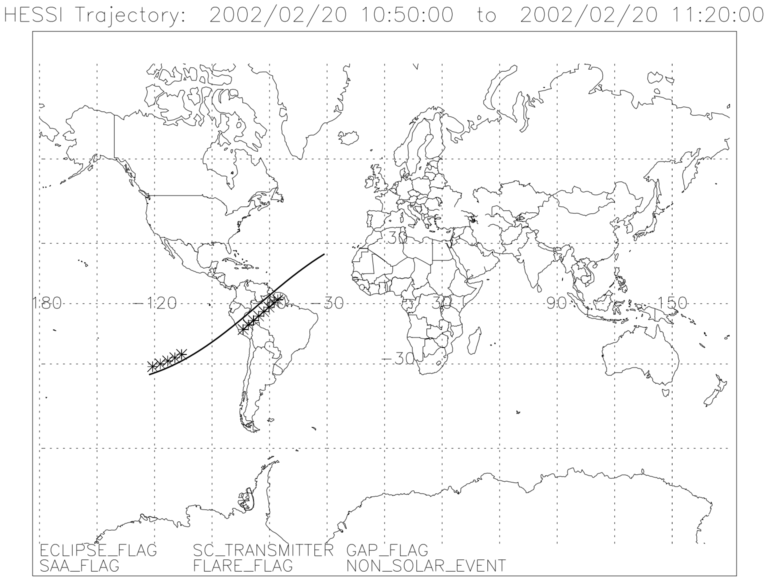

Figure 5. For each event, HEDC has a projection of RHESSI's subpoint on a mercatorian view of the Earth. A list of Observing Summary flags is shown at the bottom of the plot. Those are the flags that are being checked for and displayed with stars along the trajectory. This normally color-coded map is available with the CD-ROM material accompanying this volume.

- full-Sun image (Figure 8), using back-projection.
- movies, i.e. series of images in the following energy bands: 3-12, 25-50, 50-100 and 100-300 keV.
- 'quicklook' images and spectra (i.e. those that are included with the raw data) are also extracted and inserted in the database.
- panel of up to 5x5 images (up to 5 different time intervals, in 5 different energy bands) of the region of interest (Figure 9), using CLEAN.
- panel of 3x3 images of the region of interest, one for each RHESSI sub-collimator, using the back projection imaging algorithm (Figure 10) and photons in the 12-25 keV energy band.

Appendix B gives a full listing of data product database attributes. The time taken to generate a single 'solar flare' event and its associated data products is less than one hour for the above list of data products. More images per event will certainly be generated later on, increasing the processing time accordingly.

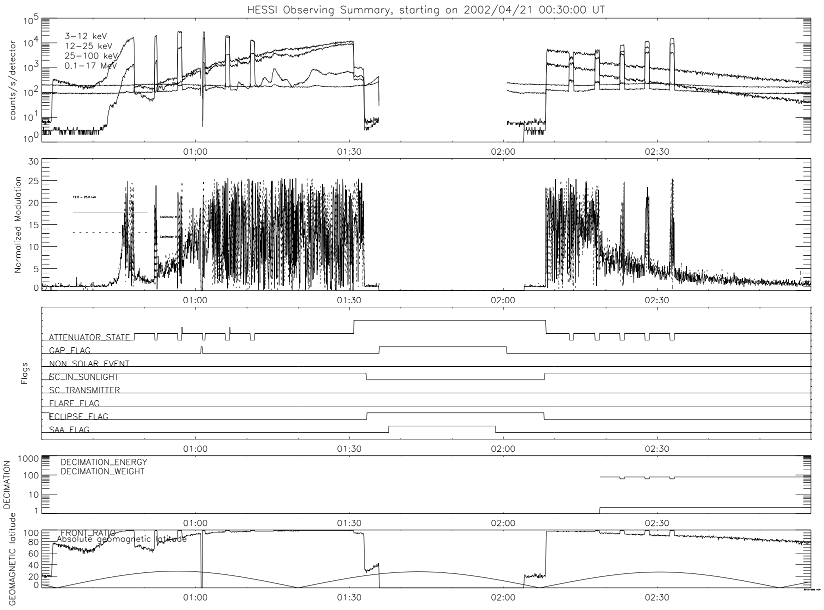

Figure 6. For each event, HEDC has an 'Observing Summary page', showing several products available in the Observing Summary, as well as RHESSI's geomagnetic latitude. The usual color-coded version is available with the CD-ROM material accompanying this volume.

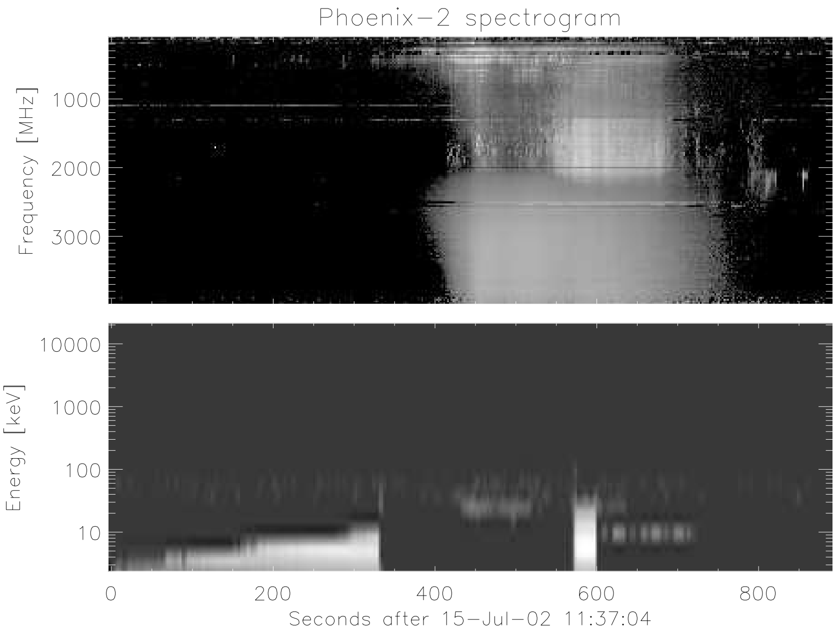

Figure 7. For each event, HEDC has a background-subtracted spectrogram from both RHESSI and Phoenix-2 radio data.

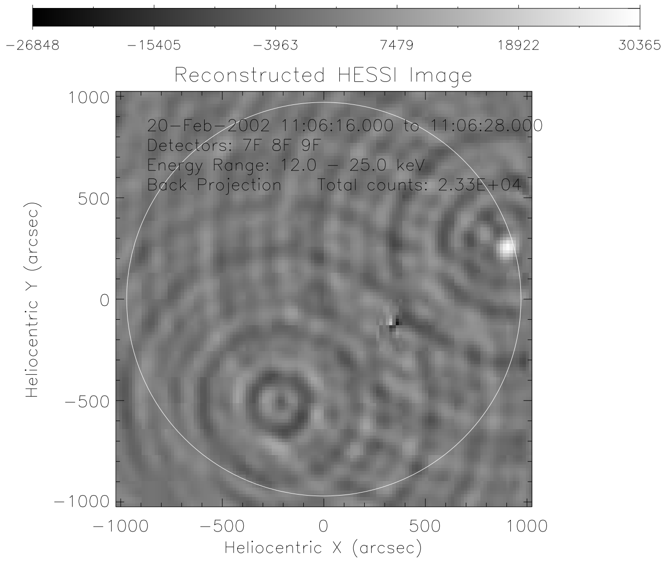

Figure 8. For each event, HEDC has a full-Sun back-projected image, at the peak of the 12-25 keV flux. Note flare position (850,280), spin axis (350,-150) and ghost image (-220,-520).

A.4. Others

RHESSI mission-long daily lightcurves in different energy bands are available through the home page.

B. Attributes used for browsing queries

Tables I and II are lists of the attributes that may be used by users to query for data on the HEDC using the Web interface's 'expert' query form. The on-line documentation provides an up-to-date listing, as well as additional details.

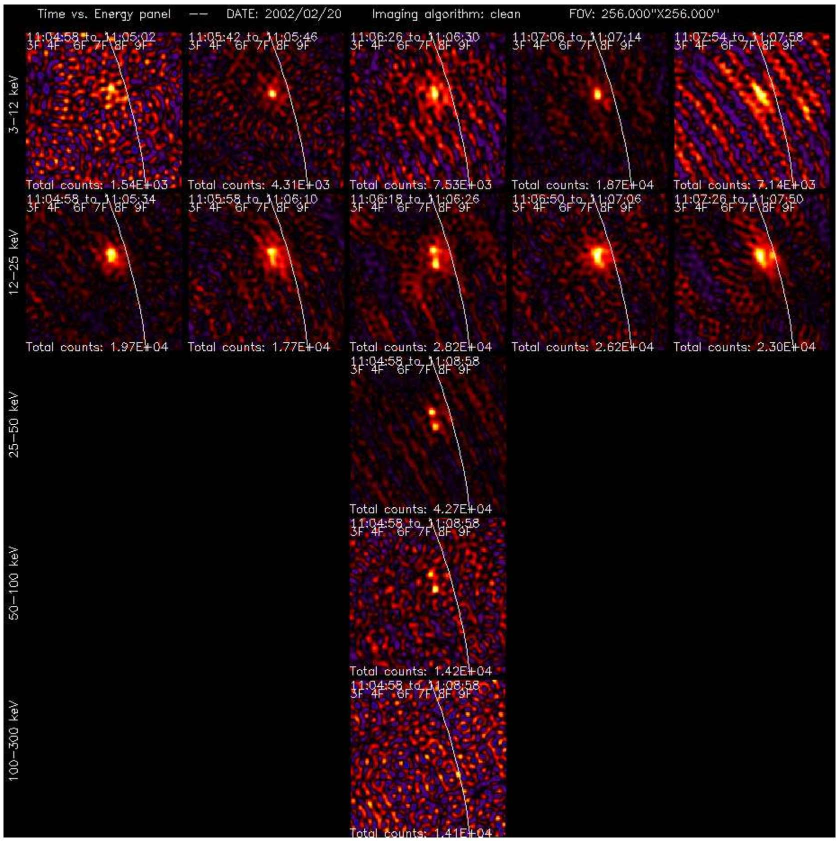

Figure 9. For each event, HEDC has a panel of CLEANed images showing the evolution of the flare in time (horizontal) and energy (vertical). Only images with a minimum number of counts are made. Hence, small flares do not necessarily have five images in every energy band.

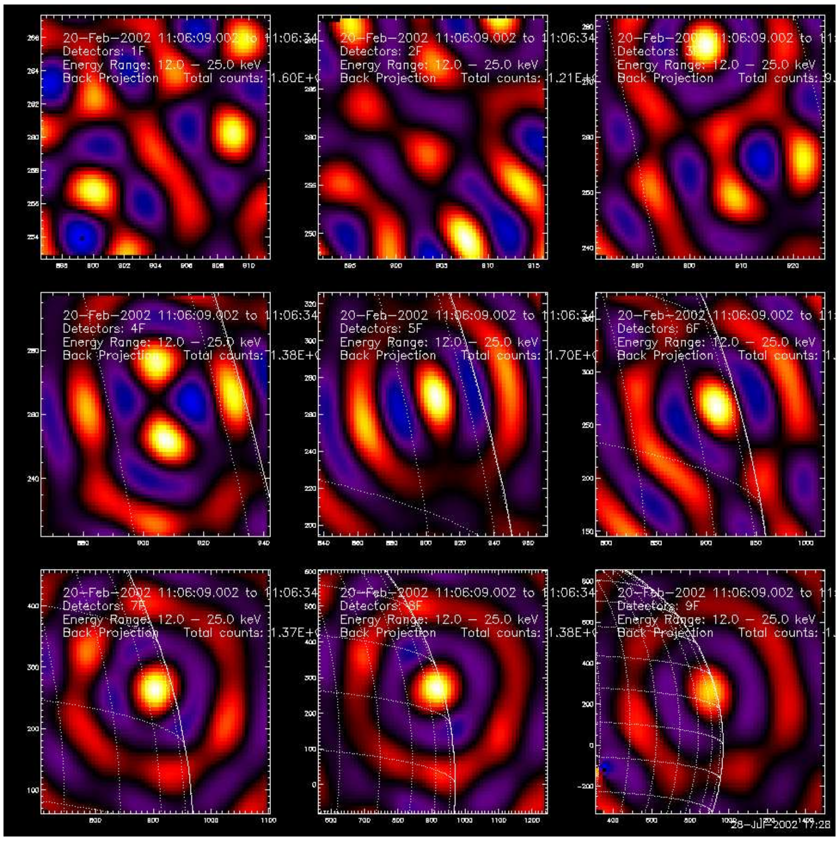

Figure 10. For each event, HEDC has a panel of back-projected images at the peak of the 12-25 keV flux, one for each sub-collimator. The field of view (FOV) increases proportionally to collimator resolution. Using the same FOV for all collimators does not allow a proper visual appreciation of each collimator's contribution to the final image.

Table I. List of query fields for HEDC events

Code	The ‘name’ of an event. Typically ‘HXS202261026’, where HX is an event made by HEDC, ‘S’ for solar flare, and ‘202261026’ for the peak time of the February 26th, 2002 10:26 flare). Another possible format is ‘hadar’, ‘hadar001’, etc. for an event generated by user ‘hadar’.
Event ID	An internal, unique ID number for each event.
Event type	‘S’ for solar flares, ‘G’ for gamma-ray bursts, ‘E’ for electron events, ‘O’ for other flares.
Flarelist	Time-concurrent flare list number for a solar flare.
Minimum energy	Lower edge of highest energy band where flare counts were seen.
Maximum energy	Upper edge of highest energy band where flare counts were seen.
Total counts	Total counts of the flare, in the 12-25 keV energy band.
Distance to sun	Solar flare’s offset from Sun center, in arcseconds.
X pos	Solar flare’s West-East offset on the Sun, in arcseconds.
Y pos	Solar flare’s North-South offset on the Sun, in arcseconds.
Creation date	Creation date of the event.
Start DATE+TIME	Date and time of the start of the flare, 12-25 keV band.
End DATE+TIME	Date and time of the end of the flare, 12-25 keV band.
Start time-of-day	Time, in seconds since midnight, of the start of the flare.
End time-of-day	Time, in seconds since midnight, of the end of the flare.
Duration	Time between flare’s start and flare’s end, in seconds.
Peak D+T (3-12keV)	Date and time of the peak of the flare, 3-12 keV band.
Peak t-o-d (3-12keV)	Peak time, in seconds since midnight, 3-12 keV band.
Total counts (3-12keV)	Total counts of the flare, in the 3-12 keV band.
Peak rate (3-12keV)	Count rate at peak time, in the 3-12 keV band.
Peak D+T (12-25keV)	Date and time of the peak of the flare, 12-25 keV band.
Peak t-o-d (12-25keV)	Peak time, in seconds since midnight, 12-25 keV band.
Total counts (12-25keV)	Total counts of the flare, 12-25 keV band.
Peak rate (12-25keV)	Count rate at peak time, 12-25 keV band.
Peak D+T (25-100keV)	Date and time of the peak of the flare, 25-100 keV band.
Peak t-o-d (25-100keV)	Peak time, in seconds since midnight, 25-100 keV band.
Total counts (25-100keV)	Total counts of the flare, 25-100 keV band.
Peak rate (25-100keV)	Count rate at peak time, in the 25-100 keV band.
Ratio 25-50/12-25	Ratio of counts in the 25-50 keV and 12-25 bands at peak time.
Source multiplicity	Number of sources in a solar flare. Not operational yet.
Active region	Where the flare occurred, as given by the flare list.
Is simulated data	0/1 or NO/YES flag.
S/C in SAA flag	0/1 or NO/YES flag. S/C stands for spacecraft (i.e. RHESSI).
S/C in night flag	0/1 or NO/YES flag. S/C stands for spacecraft (i.e. RHESSI).
Backgound rate	Background count rate. Not operational yet.
Comments	Made automatically by HEDC (e.g., highest geomagnetic latitude during an event), or by a user for a user-made event.
Reserves	unused yet.

Table II. List of query fields for data products

Code	The 'name' of a data product. For a data product associated with an HEDC-made event, the data product's code is usually the same as the HEDC event's (e.g., HXS202261026). For user-made data products, any combination of 12 characters is possible.
Product ID	An internal, unique ID number for each data product stored on HEDC.
Product type	'IM' for images, 'SP' for spectra, etc. See the online documentation for a complete listing.
Imaging algr	'BACK' for back projection, etc. See the online documentation for a complete listing.
Movie code	Most of the RHESSI images made on HEDC are meant to be viewed in sequence, i.e. they share energy bands, imagingalgorithm, etc., and differ only by their time ranges. All those images have the same movie code.
Movie frame	The order in which an image which is part of a movie appears.
Creation date	Creation date of the data product.
Start DATE+TIME	Date and time of the start of the accumulation time.
End DATE+TIME	Date and time of the end of the accumulation time.
Start time-of-day	Time, in seconds since midnight, of the start of the accumulation time for the data product.
End time-of-day	Time, in seconds since midnight, of the end of the accumulation time for the data product.
Duration	Accumulation time for the data product.
Min energy	Lower edge of the energy bands used for the data product.
Max energy	Upper edge of energy bands used for the data product.
Time resolution	Time binning for lightcurves (corresponds to LTC_TIME_RES).
Front segments used?	0/1 or NO/YES flag.
Rear segments used?	0/1 or NO/YES flag.
Subcollimator used	example: 101111100.
Distance to sun center	Angular offset (in arcseconds) of the center of an image with respect to Suncenter (image data products only).
Xpos	Angular x-offset from Suncenter of the center of an image.
Ypos	Angular y-offset from Suncenter of the center of an image.
Xdimension	Number of horizontal pixels in an image (images only).
Ydimension	number of vertical pixels in an image (images only).
Xpixel size	Horizontal size (in arcseconds) of a pixel (image data products only).
Ypixel size	Vertical size (in arcseconds) of a pixel (image data products only).
Data quality	Unused yet.
Is simulated data	0/1 or NO/YES flag.
Is background-subtracted	0/1 or NO/YES flag. Not used yet.
Other alg. params	Information on some other parameters of the data product.
Comments	Text added by HEDC or by users, for their own data products.
Reserves	Unused yet.