# The *INTEGRAL* Archive

Katja POTTSCHMIDT[1,4], Pavel BINKO[2,4], Mohamed MEHARGA[3,4], Rafik OUARED[2,4]
Roland WALTER[4,5], and Thierry COURVOISIER[4,5] on behalf of the ISDC team

[1] Max-Planck-Institut für extraterrestrische Physik
Postfach 1312, 85748 Garching, Germany
[2] SYNSPACE SA
Rue de Lyon 114, 1203 Genève, Switzerland
[3] APCO Technologies SA
Avenue de Corsier 1, 1800 Vevey, Switzerland
[4] *INTEGRAL* Science Data Centre
Chemin d'Écogia 16, 1290 Versoix, Switzerland
[5] Geneva Observatory
Chemin des Maillettes 51, 1290 Sauverny, Switzerland
Katja.Pottschmidt@obs.unige.ch

**Abstract** – The *INTEGRAL* mission is aimed at observing the sky in γ-ray light. The data obtained from the satellite are processed and archived at the *INTEGRAL* Science Data Centre (ISDC) near Geneva. In this paper, we present an overview of the available data types and their organization at the ISDC. The *INTEGRAL* archive is introduced, providing a description of its main components, i.e., the data repository, the database, and the data access and distribution mechanisms. Finally, we report on first steps in order to connect to other astronomical archive services.

## 1 – Introduction

While the main objective of this symposium is to discuss how to go beyond the organized archiving of measured data and add additional value to the measured data, the former is the basis of the latter and should already contain the seeds to facilitate interoperability and long term preservation. Therefore, the *INTEGRAL* data archive is described in the context of the "current systems and services" session of this symposium in oder to give an example of organized archiving for the case of highly complex high energy astronomy data. However, the *INTEGRAL* archive is also already including approaches which allow to manage and preserve additional information, like the grouping principle, the possibility to hold multiple revisions, and the inclusion of all processing levels from raw data to high level results. The archive is presented in sections 3 to 5. We also briefly introduce the *INTEGRAL* mission (section 1.1) and its data center (section 1.2) and describe the nature of the available data (section 2).

### 1.1 The *INTEGRAL* Mission

At the time of this workshop, ESA's International Gamma-Ray Astrophysical Laboratory – *INTEGRAL* – mission, launch date 2002 October 17, shall already be delivering new data showing the sky in γ-ray light. The prime astronomical targets in this wavelength regime are γ-ray bursts (GRBs), diffuse emission of our own galaxy, binaries containing an accreting black hole or neutron star, as well as Active Galactic Nuclei (AGN). This list implies that a wide variety of scientific objectives can be addressed. The *INTEGRAL* payload consists of the two main γ-ray instruments, the spectrometer SPI (20 keV–8 MeV, 16° diameter field of view [FOV]) and the imager IBIS (15 keV–10 MeV, $9° \times 9°$ FOV), as well as of two X-ray monitors JEM-X1 and JEM-X2 (3–35 keV, 4.8° diameter FOV) and the optical monitor OMC (V band, $5° \times 5°$ FOV). All instruments are co-aligned. The high energy instruments are based on the coded mask principle, therefore requiring complex and advanced data processing. One revolution of *INTEGRAL*'s orbit around the earth lasts 3 days. Many, but not all, data products are organized by revolution. A more detailed description of the *INTEGRAL* mission is given in the *INTEGRAL* AO1 manual (`ftp://astro.estec.esa.nl/pub/integral/AO/AODocB.pdf`).

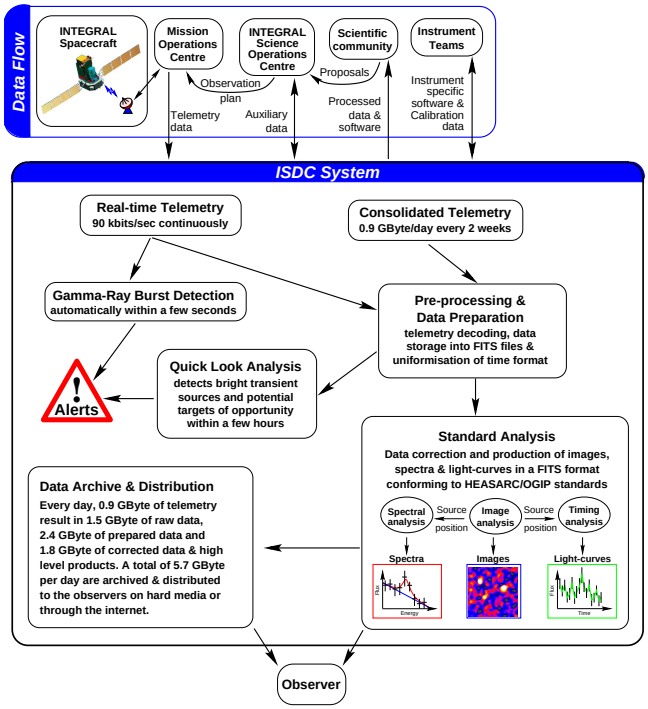


FIG. 1: Data flow during the *INTEGRAL* mission and basic components of the data processing system at the ISDC.

### 1.2 The *INTEGRAL* Science Data Centre (ISDC)

Fig. 1 illustrates the high level organization of the mission. Apart from the spacecraft itself, there are:

- the Mission Operations Centre (MOC, Darmstadt) which controls the satellite and the instruments and which receives their telemetry via the ground stations in Redu (Belgium) and Goldstone (USA)
- the *INTEGRAL* Science Operations Centre (ISOC, Noorwijk), which plans the observations according to the requirements of the scientific community
- the instrument teams (IT, several locations) which have provided the instruments themselves as well as special instrument software
- the *INTEGRAL* Science Data Centre (ISDC, Versoix)

*INTEGRAL* is an observatory which is open to the world community. The distribution of the *INTEGRAL* data to the observers is performed by the ISDC. After receiving the telemetry from MOC, the ISDC is responsible for processing the data, for evaluating their scientific content (e.g., alerts are issued in case of GRB detections), and for archiving the raw and processed data as well as results from a standard scientific analysis. The observers are provided with alert information, data, and analysis software. For a recent overview of the tasks performed by the ISDC see [2] and the references given therein.

## 2 – The Data

In order to understand the nature of the *INTEGRAL* data, several of their special properties have to be kept in mind: First, the spacecraft usually performs a dithering pattern for a given target in order to minimize systematic effects due to spatial and temporal background variations. This means that an observation is built up by a set of several neighboring pointings, so-called science windows. The need to analyze these pointings as a whole has been the main driver for the development of a concept which allows to address logically connected data sets. Furthermore, the analysis of *INTEGRAL* data involves many different types of data: e.g., those which are provided by the different instruments on several time scales, or those which are required to take the instrument response and background into account. The latter is a task which is highly complex in the case of X-ray and $\gamma$-ray detectors, especially for coded mask instruments. In both examples, a large amount of interconnected data

are involved which are stored in a large number of different files. As explained in greater detail in section 2.3, these files are organized in hierarchical "groups", which form the logical units of the *INTEGRAL* data analysis. In addition it has to be taken into account that the processing at the ISDC starts from two different deliveries of the data from each pointing (section 2.1) and that in the ISDC system each pointing is eventually represented by several processing levels (section 2.2).

## 2.1 Real-Time Data & Consolidated Data

The data from the *INTEGRAL* satellite arrive at the ISDC twice (Fig. 1). First, as a continuous stream of real-time telemetry from MOC. This data stream is directly monitored by the ISDC alert system for the detection of $\gamma$-ray bursts. A fast reaction time to GRBs ($\sim$ seconds) is a very important scientific strength of the mission. The final, consolidated telemetry is sent from MOC to the ISDC about two weeks later on CD-ROMs.
While the pre-processing and data preparation pipelines are similar in both cases, the scientific analysis is restricted to a quick look analysis for the real-time data, whereas a full standard analysis is performed for the consolidated data (Fig. 1). In this paper we concentrate on the consolidated data processing, since its results are finally stored in the *INTEGRAL* archive.

## 2.2 Amount, Format, Processing Levels

The consolidated telemetry rate arriving at the ISDC amounts to $\sim$0.9 GB/day. The ingestion of these data into the ISDC data flow proceeds via the data receipt/pre-processing components (Fig. 1). Their main task is to produce the so-called RAW data, i.e., to write all the information of the telemetry packets received from MOC into a set of FITS formatted data files. This results in $\sim$1.5 GB/day of telemetry FITS files, organized by revolution. The smallest unit of science data, the science window, typically contains a $\sim$30 min exposure. Starting from the RAW data, the pipelines running at the ISDC produce higher level data and analysis results at different processing levels. All data products are stored in FITS format (see also section 2.3). In total $\sim$5.7 GB/day are produced which are to be ingested into the archive.
Apart from the RAW data, there are two main processing levels, leading to the so-called prepared (PRP), corrected (COR), and result (RES) data. Prepared data which are organized by science window as well as by revolution are produced. The tasks performed by the associated pipelines include, e.g.:

- applying information that does not come from the telemetry data flow
- conversion of housekeeping parameters to physical units
- updating of housekeeping parameter averages for the revolution

The corrected and result data are produced during the scientific analysis of the prepared data. Each scientific analysis consists of many sub-levels. Possible corrections/results are, e.g.:

- correction of instrumental effects
- generation of good time intervals
- dead time correction
- background model generation
- image reconstruction
- generation of spectra and lightcurves

The tools available to perform these tasks are modular. How many levels are spanned depends on the context in which the scientific analysis is performed: as quick look analysis (real-time data only), as standard analysis (via the standard analysis pipeline for consolidated data, these COR/RES data are the ones that are archived) or as off-line scientific analysis (outside of the ISDC operations system, software will be provided by the ISDC). Also, the analysis products available depend on which of the instruments is considered. In this context, one of the main structural differences is, whether the corrections are performed per science window and/or per observation group. The latter is essentially an ensemble of related science windows (see next section for an explanation of the group concept). A detailed description of the data products available for each instrument, is beyond the scope of this paper, and the reader is referred to the instrument user manuals available from the ISDC web pages (`http://isdc.unige.ch/`).

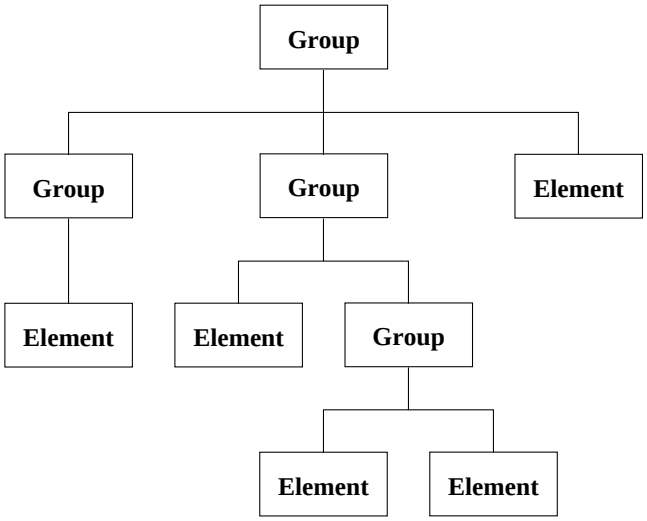


FIG. 2: Schematic view of a data object, illustrating the grouping concept.

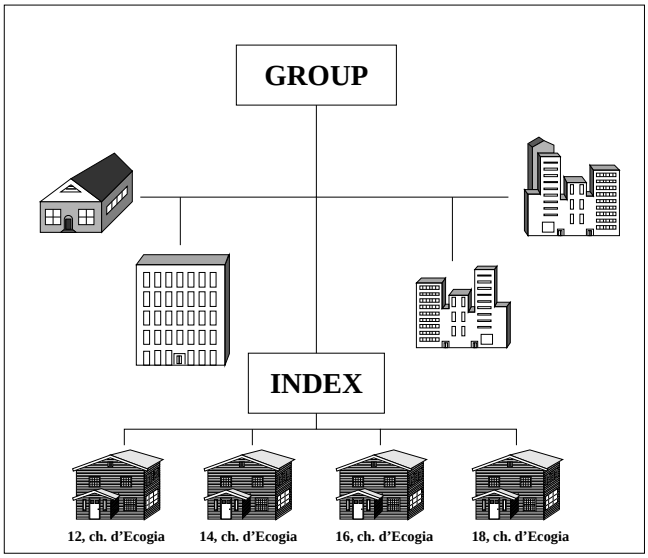


FIG. 3: Schematic of a data object containing an index (group).

## 2.3 Data Organization: Groups & Indices

A new approach is introduced by the ISDC system to provide easier and more effective access to logically connected data. The data model behind this grouping concept is the following (Fig. 2): a data object is an association of one or more data elements. A data element may be a data object itself or a base element. It may belong to many different data objects simultaneously, thus allowing data objects to share a given collection of data. The hierarchical nature of *INTEGRAL* data (which are, e.g., multi-instrument, multi-time scale, multi-processing level, and multi-pointing) make such an approach very rewarding.

The association to physical file formats is made via the FITS Hierarchical Group Convention (`http://adfwww.gsfc.nasa.gov/other/convert/group.html`): a data object corresponds to a FITS group file and a base element corresponds to a FITS Header Data Unit. The FITS group file might either contain several extensions or the locations of its members. In the case that the members of a group are of the same type, this format becomes even more efficient: in addition to the contents of a normal FITS group file, the corresponding files may also contain columns which list the values of keywords of interest for all members. Those FITS group files were first introduced by the ISDC and are called indices or index groups.

As an example, consider that each pointing consists of several data types (science, housekeeping, ... ) and is thus stored as a science window *group.* On the other hand, due to the dithering strategy employed by the

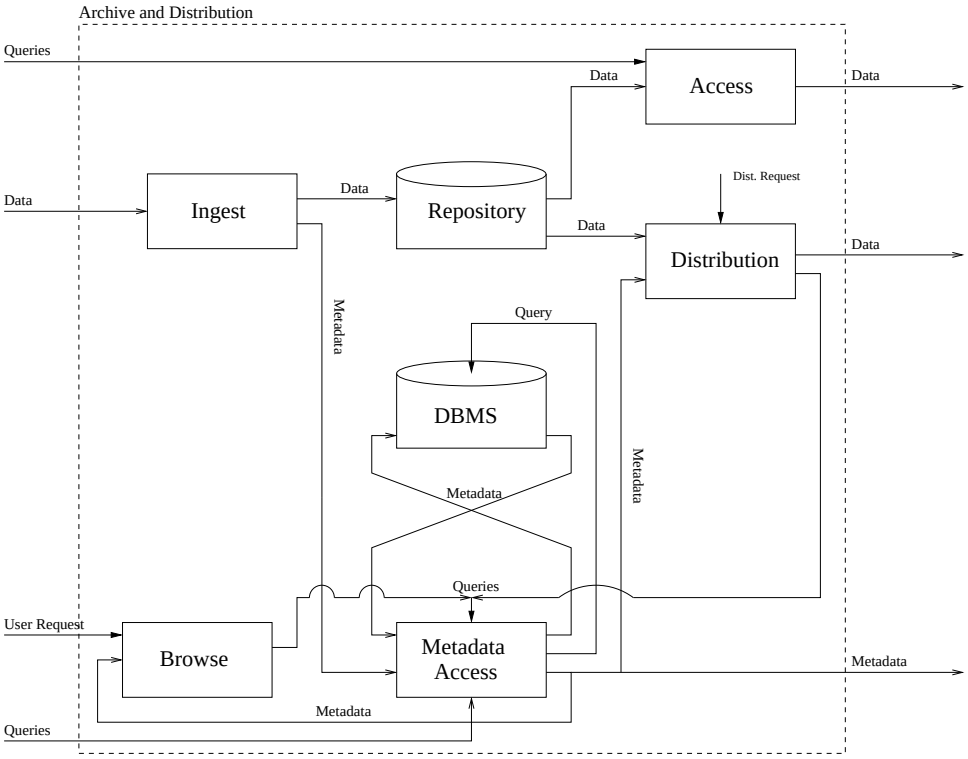


FIG. 4: Basic components of the *INTEGRAL* archive and data distribution system at the ISDC.

*INTEGRAL* spacecraft, each astronomically meaningful observation usually contains, among other things, a set of several science windows. Such an ensemble is again accessed as a group, i.e. the observation group. However, one of the members of an observation group is the associated science window group *index* (Fig. 3). The grouping concept has been implemented by the means of using the Data Access Layer library (DAL, see user documentation on the ISDC web pages for more information). This important ISDC development was built on top of the CFITSIO library. Especially, the DAL library is based on the existence of standard files templates, realized by the FITS grouping files. Naturally, the whole ISDC architecture, including the archive, is heavily influenced by the way the data are accessed, i.e., as groups.

# 3 – The *INTEGRAL* Archive System

In this section we give an overview of the *INTEGRAL* archive components, according to Fig. 4.

## 3.1 The Archive Repository

Following the grouping principle, much of the operations and analysis software of the ISDC system relies upon finding the data in a pre-determined UNIX directory structure. The *INTEGRAL* archive repository follows these rules, as well as all repositories used during operations, or those provided by the data distribution. Fig. 5 displays the structure of such an "ISDC-compliant" repository. In order to understand the management of the archive, the following elements of this structure have to be considered: below the highest level directory, which is determined by the context in which the repository is used (i.e., /isdc/arc/ for the archive), there is a separation into the FITS wrapped telemetry data which are stored by revolution (pck_X/RRRR/), and into the processed data (rev_X) [1]. The telemetry, as well as the different types of data below the processed data tree (rev_X/idx/,

[1]The ISDC system supports both, multiple archives and different data set versions, as a function of time and processing revision (denoted by the X in pck_X and rev_X). Thus the possibility is given for several repositories with the same structure to exist simultaneously. This allows to organize different phases/objectives: pre-launch testing, commissioning, operations, post operations. The data set versioning allows to ingest the same file, e.g., a science window several times, thus facilitating possible reprocessing runs.

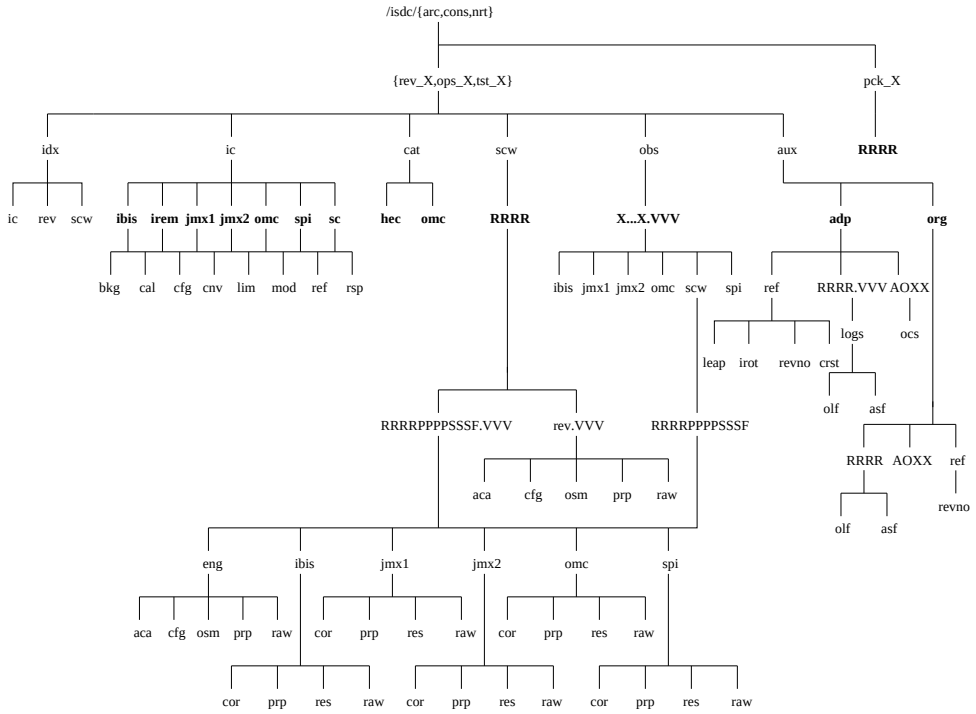


FIG. 5: The ISDC compliant directory structure which is realized in the *INTEGRAL* archive repository. The main data types in the processed data tree are: index groups (idx), instrument characteristics (ic), catalogs (cat), science window groups (scw), observation groups (obs), and auxiliary data (aux).

rev_X/ic/, rev_X/cat/, rev_X/scw/, rev_X/obs/, rev_X/aux/) are separately treated by the archive tools (Perl, C). A short description of these data types is given in the caption of Fig. 5. For the archive, different parts of this directory structure are stored on different disks, organized by mount points. In general, those mount points are located on the hierarchy level below the data types (bold directory names in Fig. 5).

## 3.2 Ingestion of Data into the Archive

The tasks performed in the context of populating the archive repository are:

- copying of the data sets to the correct place as defined by Fig. 5, the data must belong to one of the classes mentioned above (with the exception of the index groups)
- adding the version number to the file names of the files in the scw/, obs/, and aux/ (partially) directories, no files may be deleted
- validation, i.e., the extraction of meta-data (see also next section)
- indexing, i.e., the contents of idx/ is created for the ingested data

In order to accept and process ingest requests continously, there are two daemons available: the passive ingest daemon, which looks for trigger files placed into a predefined directory by an external process and creates the according entry in the ingest request queue file, and the ingest request queue daemon, which looks for new entries in the ingest request queue file. If the latter finds such entries, the ingest tool for the requested data class will be fully executed, including all points listed above.

## 3.3 The Archive Database

The archive database stores two types of data: administrative data related to the observations – e.g., proposal data and observation planning parameters – and archive repository meta-data – i.e., descriptions of the files in the archive repository and their locations. The task of the archive database is to provide users with fast access to information about the data stored in the archive repository. The standard way for the astronomical community to access the *INTEGRAL* database is via the Browse facility at the ISDC (see section 3.4.4). The browse tables

required for this are a prominent part of the database.
The operating system of the database is again UNIX and the relational database management system (DBMS or RDBMS) is provided by Oracle. The basis of the database is a conceptual data model describing the relations and dependencies of the subclasses of both the administrative data and the archive repository meta-data. In principle, the logical data model then translates this concept into tables. Apart from the directly translated database tables, the database contains two additional sets of tables: browse tables and ancillary tables. The browse tables mainly gather information about the data products accessible through Browse (e.g., describing parameters of a science window group or a observation group), while the ancillary tables store static information (e.g., the names of proposal categories, like "compact object" or "AGN").
In order to add new archive repository meta-data to the database, they have to be available in a specific ASCII format ("MDIF files"). The MDIF files are created by the ingest tools (see previous section). Again there is a daemon available, the database population daemon, which checks the predefined location for the presence of MDIF files. If such files are found, the database tables (including the browse tables) are updated by the archive database population tool. The processed MDIF files are then moved from the queue into the log area. Additional database tools (SQL, PL/SQL web applications, Perl) that are available in the ISDC system are: the database maintenance tool, the browse table maintenance tool, the database table viewer, the database consistency check tool, and the data rights manager.

## 3.4 Accessing the *INTEGRAL* Archive

### 3.4.1 Data Rights

Part of *INTEGRAL*'s scientific observation program is covering guaranteed time observations for members of the *INTEGRAL* Science Working Team (ISWT), while the remaining time is open to principal investigator (PI) guest observers through an announcement of opportunity. In both cases, however, the usual proprietary period of one year is established for the scientific data. In the archive repository, these rights are enforced via the setting of the UNIX access permissions.

### 3.4.2 Current Contents of the Archive

*INTEGRAL* will have been launched on 2002 October 17 and at the time of this workshop is foreseen to conduct its commissioning phase which is split into an instrument activation part (three weeks) and a performance verification part (PV, five weeks). The data acquired during commissioning are private, public PV pase observations are not expected to start before mid-December. However, all consolidated data at the RAW and PREP level are expected to fill the archive as soon as they are available and their meta-data will be searchable via Browse (see section 3.4.4). The higher level data are subject to extensive calibration work, before the decision will be made for these data to populate the archive. The archive will thus be extended incrementally.

### 3.4.3 Direct Access

In the ISDC system, the data in the archive repository can of course be accessed directly as far as allowed by the permissions defined above. In practice, this is especially of interest for the different projects organized in the context of the guaranteed time program: access to those private survey data of and for the ISWT is organized via UNIX group access permissions. However, these data can also be requested by ISWT members using one of the means listed in section 3.4.5.

### 3.4.4 Browse

For the general external user the ISDC provides access to the *INTEGRAL* archive by the Browse facility (`http://isdc.unige.ch/index.cgi?Archive+browse`). This web application (CGI Perl scripts) has been developed by the High Energy Astrophysics Science Archive Research Center (HEASARC) and is well established in the X-ray and $\gamma$-ray astrophysical community (for a description see `http://isdcarc.unige.`

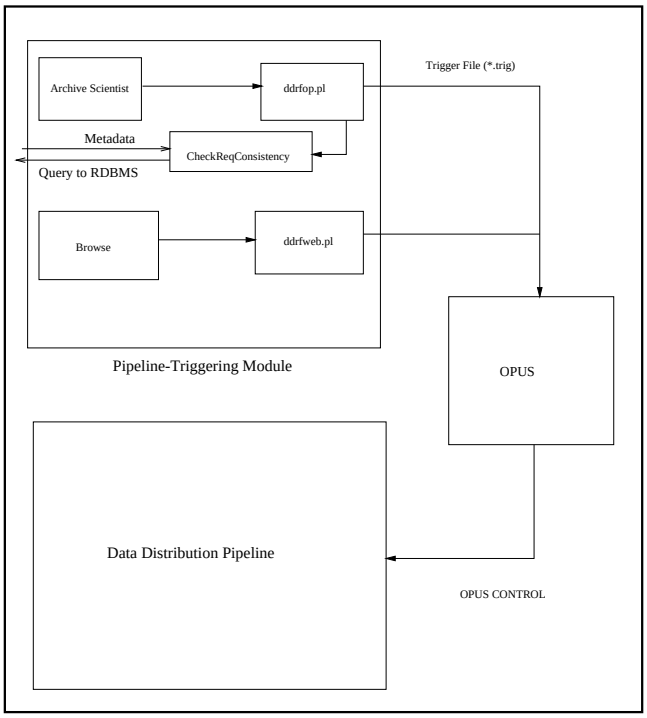


FIG. 6: Basic components of the *INTEGRAL* data distribution system.

ch/W3Browse/w3browse-help.html). For the *INTEGRAL* mission several new features were added by the ISDC, e.g.:

- selection of observation groups by multiple object coordinates
- usage of the time format ISDC Julian date (IJD), which gives the fractional number of days since 2000, January 1st (IJD=JD−2451544.5)
- handling of multiple data repositories (see section 3.1)

Using Browse, the browse catalogs associated with the archive database can be searched according to different query parameters. The primary selection criteria are given by a search radius around a given position (defined by coordinates or by an object name which is resolved by SIMBAD) and by the observation date. In principle the following catalogs can be selected and searched for each archive revision:

- auxilliary data
- instruments characteristics
- observation groups
- proposal information and observation parameters
- science windows

As output Browse provides the meta-data and administrative data stored in the browse tables for the specified data sets. If the associated data sets are public, they can be directly downloaded through the browser. In general, the query results can be used to request the data through a dedicated web application. Submitting such a request triggers the distribution pipeline which is described in the following.

### 3.4.5 Data Distribution

The main components of the *INTEGRAL* data distribution system are illustrated in Fig. 6. The system distributes private and public data to the different kinds of astronomical community users (PI guest observers, ISWT members, general public). Different methods for triggering a distribution as well as for transferring the requested data are available. Management and control is provided by an OPUS environment which, e.g, allows the processing of up to seven distribution requests simultaneously (http://www.stsci.edu/software/OPUS/). The system is again essentially realized by a set of Perl scripts. A distribution can be triggered externally by using the already mentioned web interface to submit a request (after using Browse to specify a selection). Addi-

tionally, the ISDC routinely triggers the distribution for PI guest observers as soon as their observation has been completely processed. Following each valid request the distribution pipeline creates compressed data files (tarred and gziped) containing the requested repository subsets. Depending on the method specified in the distribution request, these files can then either be transferred via FTP or on hard media. In the first case the distribution pipeline creates a randomly generated directory name under which the data files are temporarily stored in the ISDC FTP area (`http://isdc.unige.ch/index.cgi?Data+archiveftp`). In the second case the data will be written to tape and then shipped to the user. Currently hard media distribution on Digital Linear Tape (DLT, 20 GB or 35 GB) and Digital Versatile Disc (DVD, 4.5 GB) is set up. In any case the recipient is notified of the created data files by an email, which also contains a log file from the distribution pipeline. For the FTP distribution, this email also contains the location of the data file and downloading instructions.

## 4 – Data Preservation in the *INTEGRAL* Archive

On short time scales (weeks to months) the data that have been ingested into the archive repository will still be held in the ISDC system outside of the archive, thus ensuring data redundancy. On longer time scales each archive data disk will be backuped and locked when it is full. The frequency of this procedure depends on the data type, e.g., a 50 GB disk is expected to hold the science window repository branch of $\sim$10 revolutions, i.e., of 1 month. Several backup tapes of each full disk will be produced and stored in different locations.

## 5 – Interoperability

### 5.1 ESA – IESA

ESA is developing an interoperabel access to its mission archives based on a JAVA/XML browser. This approach has originally been developed for the ISO Data Archive in Vilspa and has recently been extended to also incorporate the XMM-Newton mission [1]. Its overall strategy will be presented elsewhere during this workshop (C. Arviset). It has been confirmed that this new browser can also work on the *INTEGRAL* database, namely its browse tables, without any structural modifications. However, the definition of these tables in XML will still have to be provided. Such an adaption in the context of the *INTEGRAL* Extended Scientific Archive (IESA) at ISOC is foreseen as a long term project. This early in the mission it is still in the planning stage.

### 5.2 CDS – VizieR

The Centre de Données astronomiques de Strasbourg (CDS) provides the opportunity to simultaneously query a selection from all astronomical tables in its VizeR database. Here, missions from all wavelength ranges are represented by their mission log files. Work is ongoing to also create this VizieR compatible meta-data mission log table for the public *INTEGRAL* observations on a regular basis.

### 5.3 NASA – HEASARC

The adaption of the Browse facility for *INTEGRAL* was achieved in close collaboration with the developers at HEASARC. Access to the *INTEGRAL* archive via the HEASARC Browse page is foreseen. The possibility of conducting cross correlations between the *INTEGRAL* browse catalogs and the HEASARC catalogs from other missions at the ISDC will be discussed in the future.

## References


[1] C. Arviset, M. Guainazzi, J. Hernandez, et al.. In: *New Visions of the X-Ray Universe in the XMM-Newton and Chandra* Era. Eds F. Jansen et al.. ESA Publications, 2002, in press (astro-ph/0206412).

[2] V. Beckmann on behalf of the ISDC team. In: *The Gamma-Ray Universe. Proc. XXII Moriond Astrophysics Meeting*. Eds A. Goldwurm et al.. 2002, in press (astro-ph/0206506).