Data processing pipelines and data center the X-ray spectrometer / imager STIX

Hualin Xiao[1], Shane Maloney[2], Ewan Dickson[1,4], Säm Krucker[1], Andrea Francesco Battaglia[3], László Etesi[1], and Other STIX team members

[1] University of Applied Sciences and Arts Northwestern Switzerland (FHNW), 5200 Windisch, Switzerland
e-mail: hualin.xiao@fhnw.ch
[2] Astrophysics Research Group, School of Physics, Trinity College Dublin, Dublin 2, Ireland
[3] ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland
[4] University of Graz, Universitätspl. 3, 8010 Graz, Austria

ABSTRACT

Aims. The Spectrometer/Telescope for Imaging X-rays (STIX) instrument onboard the Solar Orbiter mission launched on February 10, 2020 promises advances in the study of solar flares of various sizes. It is capable of measuring X-ray spectra from 4 to 150 keV with 1 keV resolution binned into 32 energy bins before downlinking. STIX data center is an infrastructure established at FHNW in order to process and archive STIX telemetry data, and to support the operations of the instrument. The automated data processing pipelines turn raw telemetry data into processed information and data products. Processed information and data products are achived at the data center. STIX data center provides the solar physics community various tools to visualize STIX data products.
Methods.
Results. Results.

Key words. Solar flares –Data Center – STIX data products – Data processing pipeline

1. Introduction

Solar Orbiter is a Sun-observing mission of the European Space Agency that addresses the interaction between the Sun and the heliosphere. It was lunched on Feb. 10, 2020 for a nominal mission duration of seven years and a planned extension of three years. It carries ten sets of instruments for comprehensive remote-sensing and in-situ measurements. Solar Orbiter will perform detailed measurements of the Sun as close as 0.28 AU and for the first time look at its uncharted polar regions (Forveille & Shore (2020)). Its goal is to address the center question of heliophysics "How does the Sun create and control the heliosphere?". It is designed to identify the origins and causes of the solar wind, the heliospheric magnetic field, the solar energetic particles, the transient interplanetary disturbances, and the Sun's magnetic field. This consists of the study of energetic solar phenomena like flares, solar transients, the solar wind accelerating mechanisms, and the solar dynamo principle.

The Spectrometer Telescope for Imaging X-rays (STIX) is one of the ten instruments onboard Solar Orbiter. It measures X-rays from 4 to 150 keV and takes X-ray images with a few arcsec angular resolution by using an indirect imaging technique, based on the Moiré effect . It has an angular resolution of 7 arcsec and a spectral resolution of about 1.0 keV (FWHM) at 30 keV. STIX instrument consists of 32 collimators with grids and 32 pixelated Cadmium telluride detector units called Caliste-SO. The main science objective of STIX is to study the extremely hot solar plasma and the high-energy electrons accelerated during solar flares. STIX will address the key science goals of the Solar Orbiter mission by providing information on intensity, spectrum, timing, and location of accelerated electrons near the Sun. For the details of the STIX instrument, we refer to the instrument paper Krucker, Säm et al. (2020).

During nominal science operations, STIX continuously acquires data. They are processed and compressed into different types of telemetry packets by before being transmitted to the ground. Being aware of the complexity of the data analysis and of the need to bring the data to the community, a data center has been developed at FHNW. It tasks include receiving, analyzing, archiving and distributing STIX data, and supporting STIX instrument operations. It turns raw telemetry data into processed information and data products that can be used for scientific analysis. SDC also provides various data visualization tools to the solar physics community. The purpose of this paper is to described the standard processing pipelines, data products and tools provided by the STIX data center for the solar physics community.

This paper is orgnized as follows: Section 2 briefly introduces STIX raw telemetry types, Data flow and telemetry We end with a summary in Section 13.

2. STIX raw telemetry data

STIX continuously observes high-energy events on the Sun at 4-150keV. Photons emitted by solar flares are detected with 32 subcollimators (a 12-pixel-detector behind a front and rear grid). While passing through the front and rear grids, the flares generate a modulation pattern over the 12 pixels of each detector. That pattern can then be used to reconstruct images and to do spectroscopy. Other data products include lightcurves, flare information, spectra, and background information.

The nominal telemetry budget of STIX is 50 bps. STIX is far from earth, not all data can be downloaded from the spacecraft. We have low latency data. For bulk science data have to be requested. STIX continuously collects energy deposition information from 32 detector units, the aspect system, and engineering sensors in the nominal observation mode. The collected data are first processed by the FPGA and the onboard flight software After the prompt processing, low latency telemetry data are directed to the storage in the spacecraft whearas high time resolution pixel data are written to STIX onboard archive memory for later processings. STIX transmits data to the spacecraft in the form of binary packets. STIX raw telemetry data can be be classified into four categories: housekeeping data, diagnostic data, quick-look data and science data.

In the next sections, we briefly introduce the main raw telemetry data types. For details about STIX raw telemetry data, we refer to STIX instrument paper Krucker, Säm et al. (2020).

2.1. Housekeeping data

STIX housekeeping data (HK) contain engineering data that measure temperatures, voltages, currents, status of switches, averaged signal readouts from the four aspect photodiodes, detector trigger counts and flags indicating status bits of the onboard software and the archive memoery. The data is used to monitor instrument status, and the instrument pointing. HK data are generated as long as STIX is powered on. During nominal operations, a housekeeping packet is generated every 64 seconds, which results in a daily raw telemetry of 143 kiB. All STIX sensors and the IDPU will produce housekeeping data that will be received on ground with the highest priority for monitoring instrument status and health.

2.2. Quick-look data

STIX has only one operational mode in which Low Latency data is produced, and that is nominal mode, which is the regular science mode. STIX generates four different types of quick-look (QL) data:

- Quick-look light curves, which are time series of counts of all pixels in 30 detectors (excluding background detector and the coarse flare locator) accumulated every 4 sec time resolution in five broad energy ranges (two thermal, two nonthermal and one intermediate). Quick-look light curves are generated when STIX is in nominal observation mode. The STIX lightcurve Low Latency Product is a time series of detector counts in different energy bands (units will be counts/second/cm^2). There are five configurable energy bands (for example: 4-7keV, 7-11keV, 11-16keV, 16-40keV, and 40-150keV), with a default integration time of 4 seconds. A lightcurve data point corresponds to the total counts integrated over all selected detectors and pixels for the given energy band. The data are temperature corrected on-board, and undergo minor corrections for detector and pixels, and nominal grid transmission (25 percent) on the ground. As lower energies are highly impacted by the presence of the attenuator, they will also need to be corrected. The data will not be corrected for background, live time or detector efficiency. The FITS file is in accordance with [LLFITSICD] but will also comply with in order to be usable with existing high-energy solar data file routines.
  The lightcurve data are structured as an array of five count values COUNTS (for the five energy bands) per time, where time is a relative time $RELATIVE_TIME$ since OBT_START that is time bin centered (i.e. with an integration time of 4s, $RELATIVE_TIME$ will 2-6-10..., as opposed to 0-4-8...). It is possible that that $RELATIVE_TIME$ "jumps", e.g. when the attenuator is moving in and the affected time bin becomes undetermined and is left out. Additionally, there is an energy channel reference CHANNEL, with indices into the energy axis in the second extension. Given below is an IDL-based structure notation.
- Background light curves. Background light curves are similar to the Quick-look light curves but only counts recorded by the background detector pixels are included and the integration time is 8 seconds.
- Variance data are onboard computed variance of 40 successive detector-summed count rates based on 0.1 sec integration.
- Quick-look spectra, which are energy snapshots of energy spectra (32 science energy bins) for each of the 32 detectors with 32 sec exposure time. STIX takes a snapshot of energy spectra every 1024 seconds in the nominal observation mode.
- Calibration spectra. Calibration spectra are accumulated for events from the weak onboard ^{133}Ba radioactive sources over long periods (typically 24 hours) when the solar count rate is small compared to background. The identification of such quiet periods is done autonomously, based on the presence of a TC-specified time gap between successive photons. The relevant portions of these spectra are included in the quick-look low-latency data.

2.3. Diagnostic and event data

When a failure is detected directly by STIX, the instrument's response includes sending a "dedicated" telemetry event report with appropriate diagnostic information included.

2.4. Bulk science data

Bulk science data are different combinations of summing and compression of the basic pixel data stored in STIX onboard archive memory, which can be used for spectrocopy and imaging. To cope with the limited available telemetry, they are not automatically included in the telemetry. Onboard formation of the science data is invoked by data request telecommands. Each science request can select subsets of energies, pixels and detectors. Six different science data can be generated:

- Level-0 X-ray data is the least processed data and contains uncompressed counts for the selected energies, pixels and detectors;
- Level-1 is essentially the same as Level-0 but the counts are compressed on-board before being sent to ground;
- Level-2 data are further compressed counts from the L1 pixel data, in which are summed down to 4 before compression.
- Level-3 data are visibilities, which further reduces the data by combine the the 4 summed pixel counts into a complex visibility which is also compressed.
- Level-4 data are detector summed spectrograms;
- High-time resolution aspect data.

Housekeeping data, quick-look data and calibration data are directed to the common low latency data store in SSMM in the spacecraft. The coverage, data rate and latency of different types of STIX raw telemetry data are summarized in Table 1.

Table 1. STIX raw telemetry data converage, data rate and typical reception delay at SDC.

Category	Coverage	Daily data volume	Reception Delay
Housekeeping	continous	143 kiB	hours to days
Quicklook	continous	~ 358 kiB	hours to days
Science	ground-selected only	~ MiB to ~ 10 MiB	2 to 12 weeks
Calibration	quiet sun periods	100 kiB	~ 1 day

3. Data link and data reception

SOC will host the instrument pipelines and retrieve the low latency data from MOC after downlink, passing it as input to the instrument pipelines. SOC will also post process the output of the pipelines, applying operations that will include, but not necessarily be limited to, time conversion from on-board time (OBT) to UTC and transformation of FOV parameters from instrument coordinates to an appropriate scientific coordinate system. SOC will not apply calibrations to the output of the instrument pipelines. SOC will also provide a simple web-based visualization tool for the low latency. SOC will also distribute the low latency data files via the Solar Orbiter Archive, hosted at ESAC, following the policies described in the Archive Plan

During nominal science operations, Low latency data are downlinked in the very next ground station pass regardless of orbital geometry, whereas science data are downlinked as soon as when the bandwidth permits. Solar Orbiter raw telemetry data are downlinked for eight hours during each communication period with the ground stations belonging to ESA's space tracking station network. The downloaded telemetry data are first processed by ESA's ground segment software at the Solar Orbiter mission control center. The processed data are then distributed to instrument teams by the ESA EGOS Data Dissemination System (EDDS) (Peccia (2005)) according to instrument teams' preset conditions.

STIX telemetry data arriving at STIX data center are in binary format, and they have the same format as the data generated by STIX instrument. The delay for low-latency data to arrive at STIX data center is a few hours to one or two days, while the arrival time for science data may delay weeks after being generated onboard.

In addition to the telemetry data, STIX data center also receives SPICE kernel files (Acton (1996)) from the the solar orbiter science operations center , They contains information on spacecraft ephemeris and are used for timestamp conversions.

4. Data processing pipelines

4.0.1. Raw data parsing

The telemetry reader (TM Reader) reads the input file, extracts the APID and SID, and determines the Low Latency Product to be generated. A product-dependent construction routine then transforms the telemetry content to an intermediate IDL structure for the given product (Data Extraction). Then, the intermediate IDL structure is transformed into our internal target Low Latency IDL structure (Data Product Builder). Depending on the provided product (lightcurves or flare information), minor corrections and/or normalizations are applied to the internal Low Latency IDL structure. Finally, the internal IDL product structure is written to the output NFS file share (FITS Output) using a product-dependent FITS writer (Data Product FITS Writer). This results in one file per solar orbiter packet (violet files labeled 1...N/A...N). If there are more telemetry files or packets present, the control script will continue processing, otherwise it will terminate.

Although the daily data volume of stix is small, the data structure is complex and they are highly compressed. There are more than 200 types of data packets (many of which are different types of diagnostic information). They are highly compressed and all need to be parsed. The processing includes the following processes include a decompression error map here

- Raw packet decoding Mission interface face
- SPICE kernel data is used to convert the onboard timestamp and
- Integer decompression
- Energy conversion
- Packet IO, mongodb, mongodb is used to find duplicated packets, using md5 checksum
- meta data creation

The above processing steps are done with a single program.

4.1. metadata and sysposis data creation

4.1.1. FITS file creation

5. Filename naming conventions

5.1. STIX raw data processing pipeline

5.1.1. Background data processing pipeline

Light curves measured during quiet periods of the sun are used for background estimation. Median values and of counts are computed and considered as the background in the selected time frame. They are stored in a database and used for flare identifications.

5.2. Calibration data processing pipeline

The calibration data processing pipeline is started automatically after the calibration data raw packets are being parsed and written to the NoSQL database.

As described in Ref. Krucker, Säm et al. (2020), Ba134 radioactive sources with a total activity of about 4000 Bq are placed at the front of each detector. The total activity of the radioactive source is approximately 4000 Ba. When the radioactive source decays, gamma rays are generated. These gamma rays can form peaks in the energy spectrum of the detector. As the energies of the peaks are known, and the corresponding relationship between ADC and real energy can be calibrated through the position of the peak in the energy spectrum, that is, the calibration coefficient. The figure below is a typical Ba133 gamma-ray energy spectrum measured by STIX CdTe detector. There are three obvious peaks in the energy spectrum, and they correspond to three energies of 30 keV, 35 keV and 81 keV. There are many ways to determine the position of each peak, you can use the ECC method, or use the Gaussian function to fit the left part of the peak.

ELUT and calibratino factors can be downloaded from STIX data center by using web GUIs or python APIs.

5.2.1. calibration analysis production, ELUT

6. Data request procedure

For detected flares, a script is used to create data requests. If the background subtracted peak count rate of a flare is greater 150, both L1 and spectrogram requests are created; For micro-flares, as only spectrograms are request normally as the statistics is too low to reconstruct flare images. Aspect data requests and some extra data requests are created for events of interest in the STIX operations team. For external users, data request forms can be also submitted using a web tool at the data center.

An unique ID is assigned to each data request automatically. The ID naming convention is yyddmmxxxx, in which yyddmm indicate the observation year (without century), month, date, and the last four digits indicates the serial number of the data request in the day. IDs are used to track the status of data requests, and reterieve data products from the STIX data center. The information of created data requests are stored in the NoSQL database on the same server. They are converted to instrument operation requests (IORs) after a series of checks. Requested data will arrive at STIX data center within a time frame of two weeks to three months, depending on telemetry allocations.

7. Flare processing pipeline

7.0.1. Flare identification

Quick-look light curves in the energy range 4 to 10 keV are used for solar flare identification. The flare identification procedure consists of two steps:

– Light curve smoothing. In order to filter spikes from electronics and to reduce the amount of variation due to statistics and the onboard integer compression, light curves are smoothed by using the average filtering. The
– Flare identification. Local maximums are selected from smoothed light curves. A local maximum is considered as a flare peak if the counts are exceeds 2 standard deviations above the background and the duration above the background is longer than 1 minute.

For each of identified solar flares, the information such as start time, end time, counts, background subtracted counts is stored in the STIX flare database. It is used for automatic creation of data requests for on-board archived data.

How to define a flux.

7.0.2. flare ID naming convention

7.0.3. Flare locations

7.1. Solar flare list

7.2. Flare location database

A list of products from the flare processing pipeline
* Flare list * Flare location *

7.2.1. Flare location using coarse flare data

7.2.2. Flare classification using GOES x-ray flux

8. STIX data products

8.1. L1 data products

– Raw data.
– L1 data.
– L2 data

The latest level 1 FITS IO from Shane has been integrated into the data processing pipeline on pub023 server.

Table 2. Level 1 data products

Category	Type	Naming convention	Remar
Housekeeping	hk_mini		Housk
	hk_maxi		houske
Quick-look	light curves		Quickl
	variance		varianc
	spectra		
	background monitor		
	flare location		
Calibration data			

I have recreated fits products for all old telemetry data with the upgraded SW.

The L1 fits files created by this pipeline have a different data level: L1A ('A' here means prerelease/alpha version).

The idea behind L1A data sets is to allow for quicker access to STIX data in the fits format instead of grabbing data from plots or using JSON requests,

for operations, debugging etc.

The L1A data sets can be generated within a few minutes after the arrival of a new raw telemetry file.

The differences between L1A and L1 available in Shane's ftp include:

1. Two different L1A data files may have duplicated data

2. L1A data sets are still created for incomplete packets (L1 checks for data completeness)

3. SPICE kernel data for telemetry files always arrives after one or two days later. So there may be a sub-second difference between the UTC time in fits files (same to times on web pages)

and the real time.

Shane's formal L1 release can avoid this issue if they are produced on a later date.

4. L1A contains housekeeping data

The different data-processing levels for HMI are summarized as follows, with more details available from this source: • Data at Level 0 are images that have been constructed from the raw telemetry stream. • Data at Level 1.0 are images that have been converted from Level 0, with processing including bad-pixel removal, flat-fielding, and quality assessment checks, but otherwise not having undergone any irreversible data alterations. • Data at Level 1.5 are images of the physical observables (Dopplergrams, magnetograms, and continuum images), which were constructed from the individual Level 1.0 filtergrams. • Data at Level 2 have been irrevocably filtered, time-sequence-merged, Fourier-transformed or otherwise changed from Level 1.5 data in a way that is irreversible. Level 2 products include intermediate products for later production of mission science data products, such as helioseismic inferences of solar subsurface flows

The Level0 archive contains TM which has been parsed or decommuated into readable structures but no additional external information is include:

times are not converted to UTC no calibration or conversions applied for STIX we need to decide if we decompress / combine X-ray L0 the count/trigger data at this stage or in the next level L1

copy manual, tree like, json formats name, raw value, eng value, children look-up table, to know description

estimate mongodb benchmark Mongodb benchmark, key value, index, performance

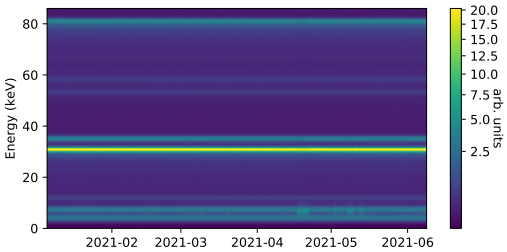

Fig. 1. Caption

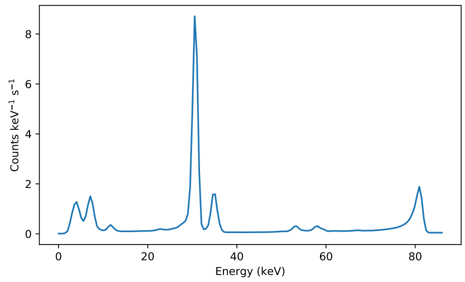

Fig. 2. Spectrogram of STIX calibration spectra

8.2. Auxilary data

9. data request principle

10. Database

10.1. Raw data packet database

10.2. Configuration database

11. Online data visualization tools

11.1. Quick-look light curve

11.2. Science data quick analysis

11.2.1. Calibration data

calibration data products https://fermi.gsfc.nasa.gov/ssc/data/access/gbm/

11.2.2. Solar Orbiter orbit viewer

When a new data file from the platform is received at the PPDC, it triggers an autonomous start of the dedicated program that decodes and interprets its contents. The binary data contain the spacecraft location, attitude, speed, and GPS timestamps with increments every half second. The GPS timestamps are converted into Unix-timestamps, where the leap seconds are also considered. After processing, the platform data are written to the ROOT format files. The data start and stop time, data processing time, input filename and ID of the output file of each processing are recorded in a dedicated database table. SPICE kernel

Updated once per day.

At the center of the Sun. It is worth mentioning that has to corrected for. This can be done by using the web tool provided at the auxiliary data center at

11.3. Automated data request scheme

scheme

12. Data access and APIs

13. Summary

References

Acton, C. H. 1996, Planetary and Space Science, 44, 65, planetary data system
Forveille, T. & Shore, S. 2020, A&A, 642, E1
Krucker, Säm, Hurford, G. J., Grimm, O., et al. 2020, A&A, 642, A15
Peccia, N. 2005, in 2005 IEEE Aerospace Conference, 3988–3995