European Grid of Solar Observations

Project n. IST-2001-32409

Title	**Unified Model for Solar Metadata**
Document number	EGSO-WP4_D1-20031009
Date	October 09, 2003
Editor	**Kevin Reardon -** ***INAF / Osservatorio Astrofisico di Arcetri***
Contributors	Silvio Giordano – *INAF/Oss. Astron. di Torino* Mauro Messerotti, Max Jurcev – *INAF/Oss. Astron. di Trieste* Bob Bentley, Nathan Ching – *UCL-MSSL* Giuseppe Severino, Ivan De Marino, Roberto Alvino – *INAF/Oss. Astron. di Capodimonte* Dave Pike – *RAL* Andre Csillagy – *UAS* Frank Hill – *NSO*
Distribution	Public

Document Version History

Version	*Date*	*Released By*	*Details*
	14 May, 2002	Antonucci	*Notes on the Classification of Solar and Heliospheric Observations*
	01 October, 2002	Reardon	Data model discussion document
0.5	25 March, 2003	Reardon	Initial Draft
0.6	22 April, 2003	Reardon	Revised Draft
0.7	29 April, 2003	Reardon	Revised Draft
0.8	08 May, 2003	Giordano	Internal Review
0.9	04 June, 2003	Reardon	Revised Draft
0.9.1	11 June, 2003	Agostini	Revised Draft
1.0	13 June, 2003	Reardon	Final Release Version
1.1.1	17 June, 2003	Agostini	Revised Release Version
1.2	11 July, 2003	Giordano	Revised Release Version
1.3.1	15 August, 2003	Reardon	Draft Revision
1.3.5	01 September, 2003	Reardon	Draft Revision
1.3.7	15 September, 2003	Reardon	Draft Revision
1.4	09 October, 2003	Reardon	Updated Release

Note: This document will continue to undergo revision to incorporate improvements or corrections to the data model as learned during the implementation phase.

Table of Contents

1. Introduction

A primary goal of the European Grid of Solar Observations is to allow for the more efficient identification and utilization of solar data from a broad range of sources. In particular, the system will provide the ability to pose queries that jointly address multiple data archives. EGSO will provide the infrastructure to interface with multiple resources, presenting them to the user in a coordinated manner.

The system will utilize the metadata describing the objects that are to be incorporated in the application. This metadata includes: a) *observational metadat*a, describing the contents of the stored datasets; b) *administrative metadata*, describing the organizations and resources known to EGSO and the means of accessing those resources; c) *derived metadata*, describing additional information extracted from observational data, in particular lists of solar features and events; d) *processing metadata*, describing the way in which the dataset was operated upon and modified. These metadata will be made available to the system through different means, both automatic and manual, and from a variety of entities. There may be duplicate metadata from multiple sources and there may be no clearly defined distinction between the metadata of different types.

To provide a coherent organization of these metadata, it has been decided to create a *data model* that incorporates the different elements known to the system. The data model must capture the required concepts from the solar physics domain, reflecting the realities and components by which scientists operate, while at the same time distilling those elements into a scheme for which a sensible implementation is possible [1]. This data model should reflect the actual concepts that play a part in the solar physics domain, without however imposing constraints based on the assumed usage of those objects. The data model should allow for flexibility and adaptation to future changes within the field. The data model will have multiple realizations, both within EGSO and for usage by solar physicists themselves.

One of these realizations will be a unified catalog of solar observations (the UOC – Unified Observation Catalog) in which the observational metadata concerning data from the whole range of solar instruments can be concisely described to facilitate searches and data mining. The UOC will be constructed through the mapping of multiple distinct representations of observational metadata from different sources onto the common terms and parameters as defined within the data model. The UOC is not intended to be a single, monolithic catalog, but rather refers to the idea of the virtual existence of the observational description for multiple instruments mapped onto a common model. The UOC may find use for a variety of purposes, and the UOC might be realized with differences in content according the requirements of a particular application.

The role of EGSO is to use the metadata available from the data providers themselves to construct the contents of the UOC. The assumption is that information describing a set of observations is presented through some stable interface and that this information can be operated upon to produce a catalog in the desired format for the UOC. The level of completeness of the metadata presented will vary greatly among the different providers. The assembly of the individual observation catalogs to provide a coherent information set is not the province of the UOC nor, strictly, of EGSO itself. The UOC will, however, provide some guidance as to what information is minimally necessary in an instrument observation catalog in order to have it reasonably incorporated into the UOC. Future observing catalogs might be produced so as to be immediately compatible with the EGSO data model and allow for a simplified incorporation in the UOC.

The primary function of the UOC will be to allow coordinated searches across multiple data sources. The goal, therefore, is not to incorporate the full description of an observation, including a complete set of instrument specific parameters, but to give priority to those components that will be of primary use in the major part of the searches to be performed by EGSO. Additional "non-standard" searches around may be enabled by the inclusion of secondary details in the UOC. In particular, the contents of the UOC should be formulated in such a way so as to avoid "false negatives", incorrectly deducing that a certain dataset does *not* match the search criteria, while "false positives", erroneously including a dataset in the list of observations matching certain criteria, is undesirable but not necessarily to be considered detrimental to the operation of EGSO. The

metadata in the UOC may also find use in the presentation or analysis of the associated data, but it cannot be assumed that such information will be sufficient in all cases. The data model will be used not only in the description of primary data, the "raw" observations of the solar atmosphere and the heliosphere, but also in the description of secondary data, a product that is the result of processing of some assembly of primary data.

We describe herein a data model for EGSO to allow an appropriate description of solar and heliospheric data according to these listed purposes. Section 2 presents a broad overview of some of the common methods by which solar data is obtained and how that may affect the definition of a data model. Section 3 discusses the metadata that is used to describe these observations and how these are currently organized into catalogs. Section 4 shows how these metadata from various sources could be used together. Section 5 then presents the approach used to define a data model for use within EGSO while Section 6 presents the data model in detail. Section 7 then touches on the different ways in which this data model might be realized or implemented. Further details about the concepts incorporated in the data model are left for the appendices.

2. Solar and Heliospheric Data Description

The primary role of a distributed system such as EGSO is to allow the sharing of primary data and secondary products generated from those data. The majority of these resources will be related to observations of the Sun, including its extended atmosphere reaching into interplanetary space. It is the goal of EGSO to construct an infrastructure capable of interoperating with the full breadth of solar and heliospheric data. While the functioning of the system itself should retain an independence from the exact content of the interchanged information, it is nonetheless informative to examine the nature and properties of the data upon which the system will operate.

In the application under development, the resources to be incorporated can be usefully divided into two primary classifications: a) remote sensing and b) *in situ*. The difference between these two types of observations is not rigidly defined, though a common description is that the latter directly sample particles and electromagnetic fields, while the former observe photons. An additional distinction, particularly important from the application point of view, is that searches for remote sensing observations are generally performed on the direction of the instrument's *pointing*, while *in situ* data are generally indexed according to the *location* of the instrument itself. This is related to the fact that *in situ* measurements of particles and fields generally provide their most precise information about the conditions local to the instrument, as opposed to remote sensing instruments, which are designed to obtain measurements of distant volumes of space.

2.1 Measurable Properties

Physical principles govern the types of information that can be carried by propagating particles or waves. Instruments, in all their varied forms, are thus limited in what exactly can be measured from the incoming waves or particles to which they are sensitive. These essentially unchanging physical properties, therefore, give a common basis on which to describe varied datasets, rather than the specific techniques used to obtain the data, which undergo rapid evolution as technologies and instrument design progresses.

An electromagnetic wave can be fully described by its propagation vector, energy, and polarization. These quantities, together with the time of arrival of the wave packet, define the limits of what can be measured from a single incoming photon. In practice, it is the average properties of an incoming ensemble of photons that is determined. Thus, we can then define four general measurable properties:

- Propagation Direction
- Energy
- Time of Arrival
- Polarization

These properties define a multi-dimensional space given by the following coordinates, including the full three dimensional representations of certain properties:

- 3 Spatial Dimensions [*x, y, z*]
- Energy [*E*]
- Time [*t*]
- Polarization [Q, U, V]

Similarly, particle measurements can be described by the physically observable parameters for baryonic matter[1]:

- Propagation Direction
- Energy / Mass
- Time of Arrival
- Charge

Direct measurements of magnetic or electric fields can provide the following information about the field at the observation location:

- Field Vector
- Field Strength
- Time of Measurement

2.2 Physical Observables

It is from the measurements of this limited set of observable quantities that a multitude of determinations of the actual physical conditions can be determined. These physically observable parameters are the true values of interest in scientific studies, as they provide the insight to the underlying processes that are responsible for the observed properties. Some of the more common physical observables include:

- Intensity / Flux
- Velocity
- Energy
- Magnetic Field
- Electric Field
- Density
- Pressure
- Composition
- Temperature
- Acoustic Power

The means by which an instrument measures the incoming waves or particles determines what observable properties can be measured. Multiple instruments may use different means to determine the same physical parameter. The physical observables for a given instrument are also determined by the techniques used in analyzing the data, and hence may evolve with changing knowledge of the instrument or involved physical properties. The measurement of a physical parameter may be more or less direct, depending on the type of instrument used. For example, *in situ* instruments can directly sample the magnetic field, which remote sensing measurements may have to use more elaborate techniques to extract field strengths from the modifications of the observed radiation field.

2.3 Remote Sensing Techniques

The Sun emits a prodigious number of photons, covering the entire electromagnetic spectrum[2]. The measurement of these photons provides a means of remotely sensing the conditions present on the Sun, extending from the core out to the heliosphere. Instruments located throughout the solar system provide measurements of the arriving photons, allowing for the separation of the incoming photons into multiple bins. This sampling of the photons is performed in such a way as to provide resolution, to varying degrees, of the information in several independent coordinates: spatial, temporal, spectral, and polarization. These coordinates are defined by the underlying physical

[1] We ignore those properties, such as spin, which are not measured by any currently deployed instrumentation.

[2] We do not include neutrino measurements in our discussion at the present time, though these too may constitute a remote sensing measurement of the solar conditions.

description of electromagnetic waves and are essentially invariant within the current paradigm. The means of performing the sampling however is essentially only limited by the imagination of individual instrument designers. It is possible to identify several common types of instrumentation, driven by the available hardware or the determined efficiencies in obtaining certain information.

The most common form of instrument uses focusing elements to form an image of the Sun as projected on the plane of the sky as seen from the observing location. The image may cover the entire solar disk (i.e. *full-disk*), a limited region on the solar disk, or portions of the corona extending beyond the visible disk (or some combination of these). Spectral information is determined by using filters or other elements to limit the energies of the photons that are detected. These images are generally recorded with two-dimensional detectors, such as charge-couple devices (CCD). The temporal coverage for a single image is defined by the beginning and end of the exposure during which the detector was exposed to the incoming photons. In the longer and shorter wavelengths, where detectors often only have a single resolving element, it is common to reconstruct images of the observed source from time series of unresolved measurements.

Alternately, while continuing to use two-dimensional detectors, it is trade the sampling in one spatial direction for a multiple measurements in the spectral dimension. Called spectrography, it is achieved by using dispersive elements to separate photons by energy onto different locations on the detector. Most often, a slit is used to sample a number of spatial points in a direction orthogonal to that of the dispersion, in this way creating a two-dimensional array in λ, x.

Additionally, it is possible to shift the slit location, generally in a direction perpendicular to the slit orientation (i.e. y), to make measurements with a fuller spatial coverage. Such techniques, referred to generally as *rastering*, employ repeated measurements made with controlled semi-continuous changes in the positioning along one or more primary axes. This can be a spatial translation, as described above, or the tuning in spectral response to provide more extended coverage in the energy scale. Even a time-series, a repeated series of measurements at regular temporal intervals could be considered a raster in the temporal dimension. Many rastering schemes, actually mix the temporal dimension with the other coordinates over which the raster is being performed (i.e. raster positions are not measured simultaneously), though this need not always be the case (e.g. a multi-slit spectrograph).

While these are some of the more common techniques used for remote-sensing measurements in solar physics, there many other sampling techniques that fall outside of these "classical" methods. Slitless spectrographs may form two-dimensional images that convolve spectral and spatial information. The use of rotating modulators can permit the reconstruction of a two-dimensional image from the signal recorded by a single integrated detector. Two-dimensional detectors with the inherent ability to discriminate among photons of different energies will generate datasets different than those currently in use. Pure photon counting instruments will measure the properties of each individual photon, providing much greater detail about the incident flux.

2.4 In Situ Techniques

As opposed to remote-sensing measurements, *in situ* observations provide information primarily about the conditions local to the instrument itself. These localized measurements provide some insight into the larger structures that strongly govern the local conditions. Different regions of the solar system are governed by the magnetic fields of either the Sun or planets. Since the current project is essentially interested in the solar conditions, we will consider *in situ* data obtained in the heliosphere, i.e. outside of planetary magnetospheres, since it is here that these measurements can be directly correlated with each other and remote-sensing observations. This provides a restricted, but coherent, set of instruments and data, including those instruments on the Ulysses, STEREO and Solar Orbiter spacecraft as well as many explicitly heliospheric monitoring missions.

With this assumption, therefore, the *in situ* data are generally related to the solar wind and the electric and magnetic fields in the region through which it is flowing. We must also consider energetic particles passing through the region that result from solar activity, though these may be more strictly classified as remote sensing observations. The quantities being measured are the composition, density, temperature, velocity and direction of the solar wind, orientation and strength

of the ambient electric and magnetic field, and flux and energy of energetic particles. Many of the measurements made are time series of quantities with one or more dimensions. The instrument may also be moving through space during the series of observations, producing a spatial scan similar to the rasters described in the remote sensing discussion above. The phenomena being observed may change more or less slowly than the spacecraft's motion.

Thus, the instantaneous location of the instrument is of critical importance. The description of the spacecraft's orbit is generally more complex than the description of an earth-bound observatory (whose location is given relative to the Earth's well known orbit). Some remote-sensing observations (e.g. helioseismology) have already begun to take into account precision measurements of the observatory location. However, these same generalized descriptions of spacecraft location used for *in situ* data will be similarly critical for remote sensing instruments such as STEREO or Solar Orbiter which will travel far from the Sun-Earth line.

2.5 Common Data Description Concepts

There are several common concepts that are found in both remote sensing and *in situ* measurements. These help better understand the general types of data being described.

2.5.1 Time Series

A series of observations obtain with a fixed instrumental setup, one of whose primary goals is to measure the temporal variation of the behaviour of the solar atmosphere or heliosphere, can be considered a *time series*. A time series can be obtained with no spatial resolution, integrating over the entire solar disk and corona. A series of repeated observations with spatial or spectral information may also be a time series. A time series is generally obtained with a regular spacing between subsequent observations.

2.5.2 Observable Range

Any observation will only cover some finite range in each of the primary coordinates axes. The extent of this range can be defined by the extremes of the region covered. Ranges with well-defined start and end points can be well described in this manner. The starting and ending of an exposure time or the limits of a rectangular field of view are examples of ranges of this type.

However, many ranges do not have abrupt cutoffs, but rather have a variable distribution within the defined extremes. For example, a spectral filter may have a changing transmission curve within the overall extent of its transmission curve. Such a distribution means that the entire range was not equally observed, which may be of importance in the utilization of the data. Other examples include images that have diaphragms or optical constraints that mask of portions of the overall field. It may be valuable to users to have information on the distributions applicable to different datasets in order to better understand the information truly contained in the recorded data.

2.5.3 Sampling

The sampling describes exactly how the measurements obtained were arrayed in the different coordinates axes. One aspect of the sampling is how the measurements were spaced one from another. In some cases this spacing is quite regular and can be described with a limited number of parameters. This is the case, for example, with a rectangular CCD array that has its pixels laid out in a fixed grid and can be well described by the number and size of the pixels. Another aspect of the sampling is how much of the observed range may have been covered by the obtained measurements. Again, in the case of a CCD, the coverage is generally assumed to be complete for the observed region (though mosaiced CCD's do not usually achieve complete coverage).

However, the sampling along any of these coordinates need not be complete nor regular. The time step between observations in a time series may vary and there are often periods between one measurement and another during which photons or particles are not being collected. Generally, for the sake of efficiency or instrumental limitations, rasters do not fully sample the scanned axis, nor need the steps be equally spaced in any representation of that coordinate. The sampling may be a result of the physical construction of an instrument and its optical elements, or may be created *post-facto* by the data acquisition or compression system

2.6 Parameter Space Description

As stated above, remote observations of the Sun are based on the measurement of the characteristics of the photons received at a detector. The primary observables for each incoming photon (observables) are:

- Energy [E]
- Spatial Dimensions [x, y, z]
- Time [t]
- Polarization [Q, U, V]

In addition, for *in situ* measurements, we can add the following observables for particle measurements:

- Charge [p]
- Mass [m]

For direct magnetic or electric field measurements, we also have a measure of the three-dimensional field vector, giving orientation and field strength.

- Field Vector [θ, ϕ, r]

These observables provide a finite limit on the number of parameters needed to effectively describe the essential content of a dataset. Furthermore, these parameters are often among the primary terms used in performing searches for identifying datasets of interest.

We can use these parameters to define a series of coordinate axes which describe a complete multi-dimensional space. Any observation must necessarily lie at some point in this encompassing space. Indeed, since each observation much sample some finite range within this space, we can define an observation by a volume that encloses the observation portions of this parameter coordinate system. This can be most straightforwardly by defining the interval, given as starting and ending ranges (E_{min}, E_{max}, x_{min}, x_{max}, t_{min}, t_{max}, *etc.*), which a particular observation covers. Putting boundaries on this "observed volume" immediately classifies the data among commonly searched parameters such as the wavelength, space and time of observations.

However, a classification based simply on enclosed volume is not enough since each instrument performs a different sampling of that volume. These differences are important in discriminating among datasets and determining their relative suitability for a given requirement. Therefore to describe in more detail a given dataset, we define the following additional parameters:

- Number of Samples [N]
- Coverage Factor [F]
- Regularity [R]

We assume any data acquisition can be described as being composed of a series of one or more distinct measurements with defined dimensions. The Number of Samples, or number of elements, gives the number of distinct measurements made along a given axis, spatial, temporal, spectral (e.g. N_λ, N_x, N_y, *etc.*). These measurements may not fully sample the interval (e.g. $\boldsymbol{x}_{min}$, $\boldsymbol{x}_{max}$) defined for the relative coordinate axis. The Coverage Factor therefore defines what percentage of the observed range was actually sampled during the observations. This may define the ratio of time actually spent collecting photons during a time series to the total time period of the observations, for example. Finally, the spacing of a series of observations along a certain axis may not be evenly spaced, due to instrumental constraints or observational needs. The Regularity parameter quantifies how evenly spaced the samples are within the observed range. Datasets constructed from helioseismology networks or satellite observations might be obtained with a more regular cadence than observations from single earth-based observatories with day-night cycles or weather interruptions. Together, these three parameters can provide a coherent description of a broad range of heterogeneous datasets. Such terms may not completely describe the detailed structure of the data, but it is rather formulated to facilitate searches and comparison across multiple data sources.

This definition of the parameter space volume covered by an instrument is sufficiently flexible

to be used in different places throughout the EGSO infrastructure. This format is equally capable of representing single observations (i.e. a single image, spectrum, etc.), as well as long time series of repeated observations. The parameter space description could be used in the resource registry to give a broad definition of the capabilities and coverage provided by a given instrument. It could be utilized in searches to identify datasets matching certain temporal or spatial criteria. This same statistical description could also be used to describe data presented in other coordinate systems, such as the

- Temporal frequency [□]
- Spherical harmonic degree & order [ℓ, *m, n*]
- Spatial wavenumber [k_x, k_y, k_z]

coordinate systems used in helioseismology. Again, this would provide a description of the information content of the volume enclosed in this alternate multidimensional space.

2.7 Data Processing

The data obtained by an instrument are often not immediately usable in their raw form. Some processing might be required to remove spurious instrumental or systematic effects that would hinder the basic utilization of the data. Further processing, of differing degrees of complexity, might be needed to extract a range of physical observables from a dataset. Combination of data from multiple observing times or different instruments might also be performed to extract additional information. All these possible steps define a spectrum of processing steps that may be applied to a dataset.

2.7.1 Data Calibration

The raw data obtained by an instrument often needs to undergo calibration in order to convert the raw counts measured by the detector into more physically relevant parameters. The calibration may be carried out in a variety of steps and some algorithms may be generally applicable to multiple datasets, while others are specific to a particular instrument. The calibration may be carried to various levels of reduction, producing products tailored to different purposes. These may be given explicit labels, such as *Level 0*, *Level 1*, etc., or may be differentiated as *Raw*, *Uncalibrated*, *Calibrated*. It may be important to know which software versions and processing steps were applied to generate a particular calibrated dataset. For some datasets no defined calibration procedures are avail able or foreseen.

The data providers may or may not make the data available in a calibrated form. Often, particularly in solar physics, data are stored uncalibrated or partially calibrated, with the final calibration being performed under the users' control. The providers may provide the capability to perform user requested calibration on demand, or they may provide a software package that allows the user to perform the reduction independently of the provider.

2.7.2 Secondary Data Products

The data descriptions to this point have concentrated mostly on primary data products, the essentially unelaborated data as obtained by a single instrument. However, many tasks in solar physics are more interested in secondary data products that are the result of further processing or combination of multiple datasets. Such secondary datasets may provide information on additional physical parameters derived from the observed data, or they may map existing datasets into new coordinate axes. Some examples of the former type of data products are vector magnetograms, dopplergrams, filter-ratio temperature maps, or abundance measurements. An example of data products that present the data in alternate coordinates are the ℓ–□ diagrams used in helioseismological studies. Secondary data products tend to be less connected to the specific instrumental parameters and more related to general physical conditions.

3. Observational Metadata and Catalogs

The data described above are stored as arrays (both analog and digital) of information. The access, understanding, and utilization of these arrays generally require external information that describes their contents, called metadata.

3.1 Observational Metadata

In the field of observational solar physics, data descriptions play a crucial role in the identification of data of interest for a chosen research topic. These descriptions contain information that describe the means by which the data were acquired, the coverage achieved, and other details that record the motivation and applicability of the data. These *observational metadata* may be of interest both for the identification of datasets of interest as well as for the reduction and analysis of those same data.

The observational metadata may be divided into two different classifications: a) *content metadata*, describing the bytes contained in the data, including, for example, the instrument utilized, parameters describing the observation, principal investigator, scientific program, etc.; b) *curation metadata*, that describe how those data are stored and accessed, such as file format, byte order, location, or access rights.

The observational metadata may be intended for human interpretation or for automatic processing. This may result in conflicts between more complete, but perhaps less intuitive, formalisms for describing data that may sacrifice human readability for explicit correctness. Some descriptions are not complete, relying on assumed external knowledge about the solar physics domain in order to guide their interpretation.

The common practice within solar and heliospheric physics is to store the data and an essential portion of their associated metadata together in the same physical file, through a structured header that organizes metadata describing the enclosed data (e.g. FITS, CDF). The header primarily contains content metadata, particularly those elements which may vary among observations. Other metadata may be stored externally or incorporated in related software analysis packages. The storage of data with descriptive headers (in a human readable and/or well document format) is also considered to be the means by which the data can remain usable by researchers in the future.

3.2 Observation Catalogs

Some portion of the observational metadata may be collected into organized presentations that are often called *observation catalogs*, though there is no shared understanding of what the content or extent of such catalogs should be. Observation Catalogs are generally considered to be separate from the data they describe, in that they exist as files or other entities that can be accessed independently from the data themselves. This distinction is motivated by the different methods used for interacting with the metadata as opposed to the data, as well as the often great difference in data volumes between the metadata and data.

Catalogs may have a variety of scopes, but they are generally constructed to permit searches within the overall set of data. These searches use the information stored in the catalogs to identify those data that match a certain set of input parameters targeted towards identifying data of scientific usefulness, usually based on correlations with other events (these “events” may be, for example, solar features or observations with other instruments).

The catalogs used for searches generally contain a set of metadata that was defined based on the types of searches that were envisioned or desired by the producer (i.e. the instrument team or principal investigator). These included metadata may describe the regions, spatial, spectral and temporal, covered by the observations, as well as additional information about the motivation behind the observation (e.g. solar features present, observer, etc.). This information may have no use other than for identifying data of interest. For example, the information about the observation campaign under which a dataset was obtained might be of significant interest to a user in locating a

dataset, but would not generally impact on how those data would be subsequently calibrated and analyzed. Since the goal of the data searches is to identify data relevant to a specific research project, the information catalogued is often oriented to describing the data in terms of its scientific content. Additionally, the catalogs may also maintain the curation metadata needed to facilitate access to the identified datasets. Some catalogs are also directly used during the data analysis process itself to provide input needed in the calibration of the data.

The observation catalogs often do not individually list each acquisition (i.e. a shutter opening or detector readout), but rather contain summarized information describing some coherent set of measurements. This can be done to reduce the size of the catalogs, or to describe common or necessary groupings of data. For example, each individual observation in a raster may not be individually cataloged, but rather the overall coverage of the full collection of observations may be used to collectively describe the scan. The logic by which series of measurements are grouped is highly dependent on the instrument itself and what the producer determines is the smallest observational unit of independent value. Therefore, the amount of detail found in different catalogs is variable, with some containing observational metadata about each image, while others only provide daily summaries of the data obtained. Such variability complicates the correlation of the contents from multiple observations catalogs.

3.3 Limitations of Existing Catalogs

Catalogs are provided by numerous groups, each responding to imposed constraints or perceived goals of the catalog generation effort, resulting in a broad range of content, usability, and completeness. While each catalog may be valid or sufficient within its own scope, any attempt to combine different catalogs in a coherent and automatic manner exposes the differing interpretations of the task by data providers. It will be necessary for EGSO to cope with the heterogeneous collection of existing catalogs in order to access the information necessary for its successful functioning.

The construction, or not, of an observation catalog for an instrument is dictated by perceived purposes, resource limitations, and political considerations. Most critically, this manifests itself as an enormous variation in the contents of the catalogs, which may be conditioned by the complexity of the associated instrument, by the quantity of information recorded by the control system, or by decisions concerning the relative importance of intra-project and external use. Similarly, the amount of detail about a particular observation, whether listed to the level of the single data acquisition or presented in a more condensed form covering a series of repeated acquisitions, is often related to the mode of instrument operation and the goal of the catalog.

More problematic are those instruments for which a formal catalog is not accessible through a structured interface. A public catalog may not have been deemed necessary, or only stored in some "offline" format, such a hand written log. In some cases, the pertinent information to identify an observation may be encoded in the filename of the data file or the directory structure where the data are stored. Alternatively, information may be available from external files or from the headers of the files themselves.

While these are important issues to be addressed on both technical and social levels, they are not within the scope of the current document, nor can it be expected that a data model can independently solve these problems. The definition of the data model will assume that all resources to be incorporated into the UOC present themselves with a coherent catalog containing at minimum a defined subset of "sufficient" information. The generation of the necessary catalog, where not already available, may involve both efforts by the provider to correct or complete their catalog, as well as assistance from within EGSO to assist in the proper description of the catalog content. This process is logically separate from the definition of a mapping of an individual catalog onto a common framework. However, the data model must maintain a level of flexibility in describing the solar and heliospheric data to take into account unavailable information or varying amount of detail among catalogs.

3.4 Derived Metadata Catalogs

In addition to the observation catalogs, there are catalogs which attempt to compile information extracted from the observations in order describe actual occurrences in the solar atmosphere. The metadata derived from the contents of the primary or secondary data are called *derived metadata*. These catalogs can be generated both through manual examination of the data as well as through automatic processing utilizing various techniques. The most common of these derived metadata are the lists of observed solar events or features.

A solar event or feature is an *occurrence* that takes place in the solar atmosphere or extended heliosphere. Such occurrences are generally deviations from some "mean" conditions (where the so-called mean may be taken over different temporal scales). Such deviations are interesting because they indicate the presence of additional physical forces acting on the solar plasma. The lifetimes of such influence may have a large range of time scales and spatial dimensions. There are general *classes of occurrenc*e that provide groupings of features of events of similar types.

Extracted information about individual occurrences is compiled into *solar event* or *solar feature catalogs*. Multiple catalogs may exist for a given class of occurrence, each utilizing a different definition of the required characteristics for an occurrence of a certain class. The difference among definitions may come from the different observations utilized (e.g. flares observed in Ha or X-rays), the different techniques applied (e.g. visual detection versus automated extraction), or even different understandings of what constitutes occurrences of that class.

Analogously to the situation with the varied representations found among observation catalogs, the derived metadata catalogs are produced in a variety of formats with vastly different levels of detail in their content. Because of this heterogeneity, it is often difficult to make intercomparisons of the content multiple catalogs, or to combine the content of these catalogs with the observation catalogs beyond the simplest of searches.

4. Interoperability

The merging of metadata coming from multiple sources, eliminating differences in the syntactical formatting or conflicting terminology, in order that a joint use may be made of all the encoded information, is the task of ensuring *semantic interoperability*. The differences in metadata are the result of the distributed and semi-structured environment in which many data systems have been developed. The need for interoperability among data systems has always been tacitly understood within the solar physics community, and some progress has been made in this area in recent times. However, the need for a broad agreement on a common metadata layer has only now become of obvious importance, with the continued expansion of data volumes, the need to combine multiple datasets, and the development of the software tools to allow the promise of such desires to be realized.

Of the gains that have been made in interoperability in recent decades, the most notable is the widespread adoption of the Flexible Image Transport System (FITS) as a common data storage format within the astrophysical community. The FITS Standard defines both a storage format and a list of metadata elements (represented by *keywords* stored in the FITS *header*). However, those elements defined in the standard (whose goal was not, in fact, semantic interoperability) are neither complete, nor is their usage consistent among different systems. However, the FITS headers are increasingly used as the primary storage area for the metadata describing the data.

This has led to the definition of new standards and definitions for new metadata components for use within the FITS standard. One example of this is the effort on creating a more complete and flexible description of astrophysical coordinate systems within the confines of the FITS keyword namespace. This has led to the definition of the World Coordinate System (WCS), which is being adopted as a part of the FITS standard and is becoming the widely used way to express spatial or spectral coordinates in FITS headers.

Within solar physics specifically, there has also been effort to standardize usage of certain FITS

keywords or other metadata terms. The most significant recent example of this has been the list of FITS keywords generated for the data files of the Solar and Heliospheric Observatory (SOHO). This list provides an extensive list of keywords, both general and domain specific, as well as their definitions and usage. Other projects have also adopted some portions of this list for describing other instruments as well.

Another effort furthering interoperability within solar physics is the SolarSoft data analysis environment. The data access and analysis routines from multiple instruments have been provided in a common language, the Interactive Data Language (IDL), allowing, within the limitations of the system, the mapping of multiple datasets to a unified framework. This framework is more focused on data interoperability and is constrained by the limits of the software components provided by each instrument team.

A similar effort is taking place within the broader nighttime astrophysical community to define a common model and data descriptions for different aspects of the field. This effort is being led primarily by the International Virtual Observatory Alliance (IVOA). Since there may be some common terms and concepts among solar and other astrophysical fields, we are following the work being done and will incorporate appropriate model components where such commonality is both reasonable and offers a benefit in terms of reduced development times or greater interoperability.

Outside of astrophysics itself, other organizations have attempted to provide definitions and standards for common metadata in order to arrive at a common usage and facilitate interoperability. One important metadata standard is the Dublin Core, originally developed for usage in holdings description by libraries, which encompasses many very general metadata concepts (e.g. author, contact, etc.) applicable to other fields as well.

It is important to realize that interoperability is not a problem that has been created by the recent virtual observatory or distributed resource projects. These have only heightened a problem that has been long known, but that faces significant hurdles in being adequately addressed. Efforts at standardizing metadata among multiple instruments or organizations often fail to full achieve their goals. For example, the list of SOHO keywords, defined as a standard for a single space mission, did not find full take up or consistent implementation even among the mere dozen instruments on that mission. The inertia against such standards is partially social, partially systematic problems with having a single set of metadata that is valid or complete for the broad taxonomy of solar instrumentation.

4.1 Spatial Coordinate Systems

It is of obvious importance to be able to absolutely locate a recorded observation along the various coordinate axes that describe the various dimensions of the real world. This is important for understanding the physical objects observed as well as allowing multiple datasets to be converted to a common mapping. The definition of a coordinate system for solar observations is complicated by the lack of fixed surface features on the solar surface to use as reference points and the fact that the rotation of the Sun varies at different positions in the atmosphere and according to the kind of feature observed. Furthermore, the tenuous nature of the solar atmosphere may produce an ambiguity in identifying from a single observation the exact three-dimensional spatial location of a particular observed feature, Different operational and physical aspects of the instrumentation used mean that no one solution will be ideal for all uses.

Recently, Thompson (2001) has described the different coordinate systems that are utilized in solar physics for identifying positions on the solar surface. Some of the coordinate systems identified by Thompson include heliographic or Carrington coordinates, heliocentric coordinates, and helioprojective coordinate systems. The Carrington coordinate system uses a latitude, longitude, and radial distance defined by the (empirically defined) rotation axis of the Sun and a defined rotational period. The longitude is given with respect to an arbitrary starting point and successive rotations are numbered to allow observations to be located in time as well. The heliographic coordinates of an object can be given independently of the observers' positions.

The heliocentric coordinate system is also commonly used, often because it can easily map onto the two-dimensional arrays used to observe the Sun. These coordinates are given as simple offsets,

in Cartesian or radial coordinates, relative some fixed position as defined from the observer's point-of-view and neglecting projection effects. The third dimension is given by distance along the field of view and normal to the plane of the sky. Helioprojective coordinate systems are a more generalized form of the heliocentric coordinates, allowing a more precise description of the projected plane of the sky that maps onto the observations. This distinction is important for data that cover large angles on the sky, for example when observing large portions of the solar corona.

The decision of the coordinate system to use for a given observation is closely tied to the type of instrument involved and the expected use of the data. Currently, the majority of solar observations are obtained from the Earth's surface or from low earth orbit. In this case, the sampling obtained can be concisely defined by a simple heliographic or heliocentric coordinate system. As spacecraft such as STEREO or Solar Orbiter began to make images from positions well away from the Earth-Sun line, the need to precisely define the coordinate system with respect to earth-bound observers will require a more thorough description of the coordinates used. However, it is possible, knowing the position of an observatory or spacecraft, to transform the representation in one coordinate system into another more complex coordinate system that will allow more complete comparisons with other datasets.

5. Solar Data Model

The role of a data model is to provide an abstraction and simplification of the realities of the domain in which the implemented system will operate. The data model provides a common understanding of the various concepts that make up the system, an understanding upon which communications among various components can be based. The data model is not an attempt to exactly represent all information about the domain, but rather an attempt to capture the essential abstractions that provide a better understanding of the elements of a complex system.

The solar data model should represent the metadata that describe the underlying solar data. The data model must incorporate both content and curational metadata (see Sec. 3.1). Hence the "data" model refers not only to the actual solar data themselves, but, more generally, to the broader set of information that generally defines all relevant aspects of solar and heliospheric physics.

The general outline for the solar data model is presented in the following section. The discussion of individual classes and further details on model concepts are presented in Appendix B. The data model has been diagramed using the Unified Modeling Language (UML), a standard tool for such purposes. The UML data model allows translation of the model into other standard representations, such as an XML schema. The UML model includes *classes* (boxes) that represent certain real world objects or entities, which are connected via *associations* (lines) that describe how the different classes are related.

To further simplify the description of the different areas of the data model, we have divided the data model up into *packages* that group model classes related to several macro-concepts of the model. The defined packages include:

a) Metadata – the description of common metadata concepts that span multiple packages;
b) Instruments – the description of the tools that acquire data;
c) Data Production – the description of how instruments are organized to obtain data;
d) Data Sets – the description of the datasets produced;
e) Data Provision – the description of how data are made available by one or more providers;
f) Derived Metadata – the description of the Solar Feature Catalogs that describe extracted events;

The nested organization of these different packages is given in Figure1a. It is possible to further expand the contents of the different packages in order to better see the relationships among the different classes within the individual packages, as shown in Figure 1b. For the sake of clarity, this figure shows only a selection of the entire collection of classes.

The current data model attempts to describe the high level relationships among the principle concepts in the domain. The model does not address certain, albeit important, details concerning the

specific representation of those concepts. For example, the model does not presently attempt to provide complete descriptions of quantities, units, coordinate systems, physical obervables, etc. It is expected that the precise model for these objects will be developed in parallel to and informed by the implementation process.

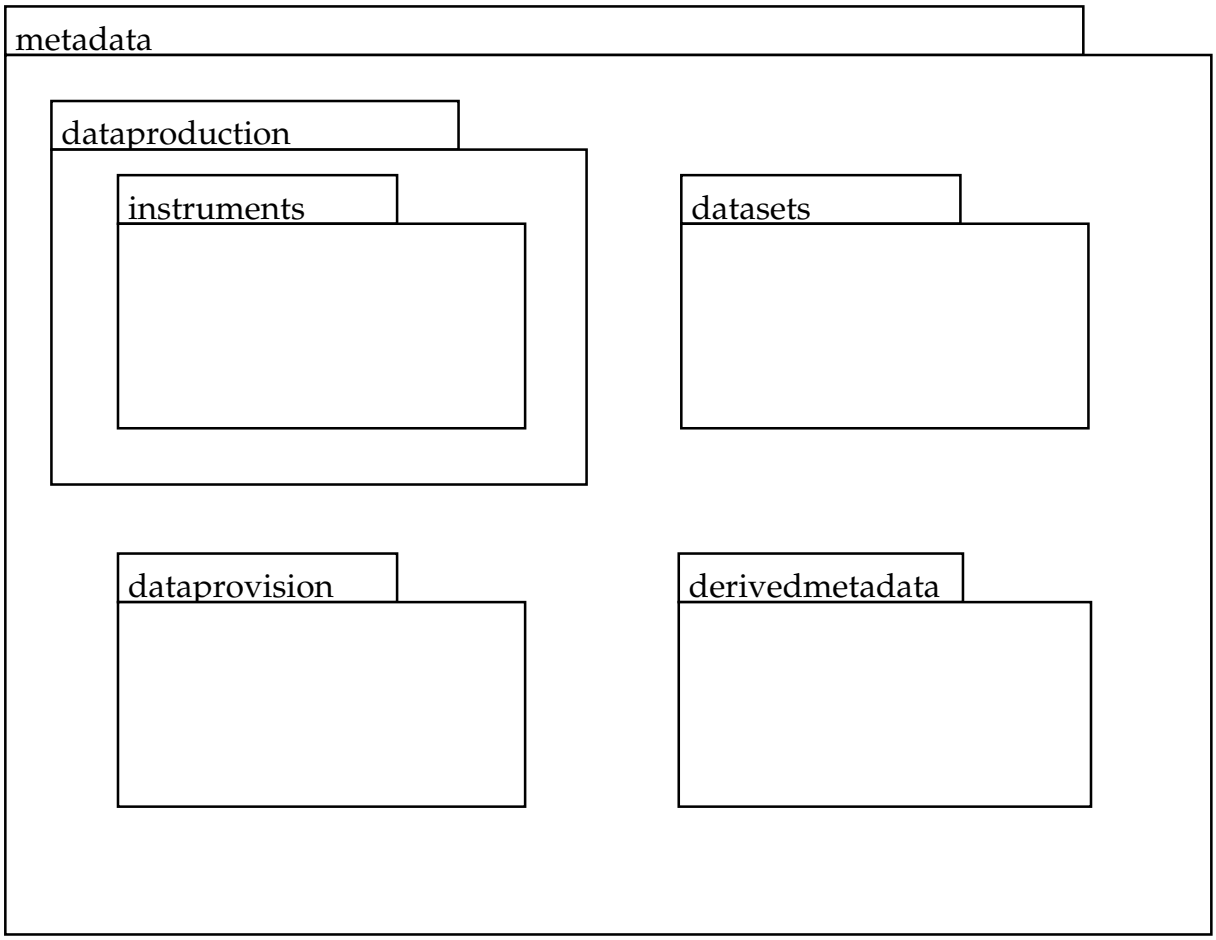


Figure 1a: Package overview of Data Model

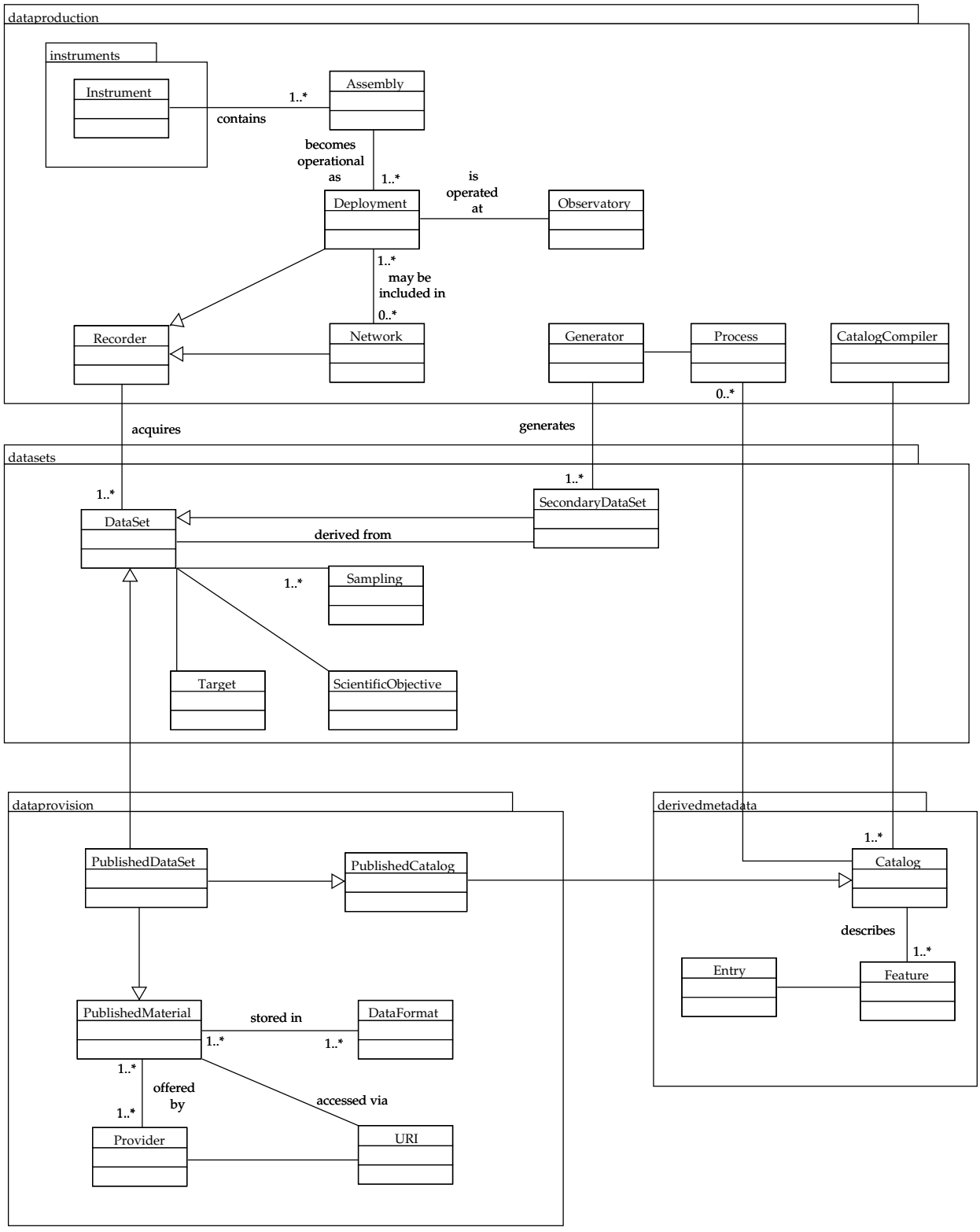


Figure 1b: Simplified overview of Data Model

5.1 Metadata Model

There are several common concepts that appear at multiple points within the overall data model. Thus we describe these concepts as part of the general metadata terms of the model. The majority of these terms are related to the placement of some element in a multi-dimensional coordinate system. Such a task is quite common throughout the model as it relates to the location of an observatory, the time span of a responsibility, the coverage of an observation, the determined position of an identified solar feature and so on.

Condensed Description:

A **CoordinateSystem** needs to be established prior to the definition of any coordinate position. A CoordinateSystem is defined by the **TimeFrame** and **SpaceFrame** that make up the reference axes for the overall system. CoordinateSystems are of one of a range of types (e.g. spherical, Cartesian, etc.) given by a **CoordinateFlavor**. Each CoordinateSystem will have one or more individual axes, identified by a **CoordinateName**. A **CoordinatePosition** is defined as a collection of positions, one for each axis, for one or more CoordinateNames. A collection of CoordinatePositions can be composed to define a **CoordinateArea** of different forms. One type of CoordinateArea is the simple **Interval**, which defines two boundary positions, each of a certain **BoundaryType**, along a single axis. One or the more common types of Interval is the **TemporalInterval**. An Interval is one of the primary components of the **Sampling** description, which defines an enclosed volume by the Interval along each axis as well as the statistical description of the distribution of sampling points within that volume. Such a sampling may be described for any CoordinateName, though **SpatialSampling**, **TemporalSampling**, and **SpectralSampling** will be among the most common.

The **Location** of an object may be described either as a fixed CoordinatePosition in a given coordinate system, or as an **Orbit** that describes the object's space-time trajectory, as defined for a specific coordinate system.

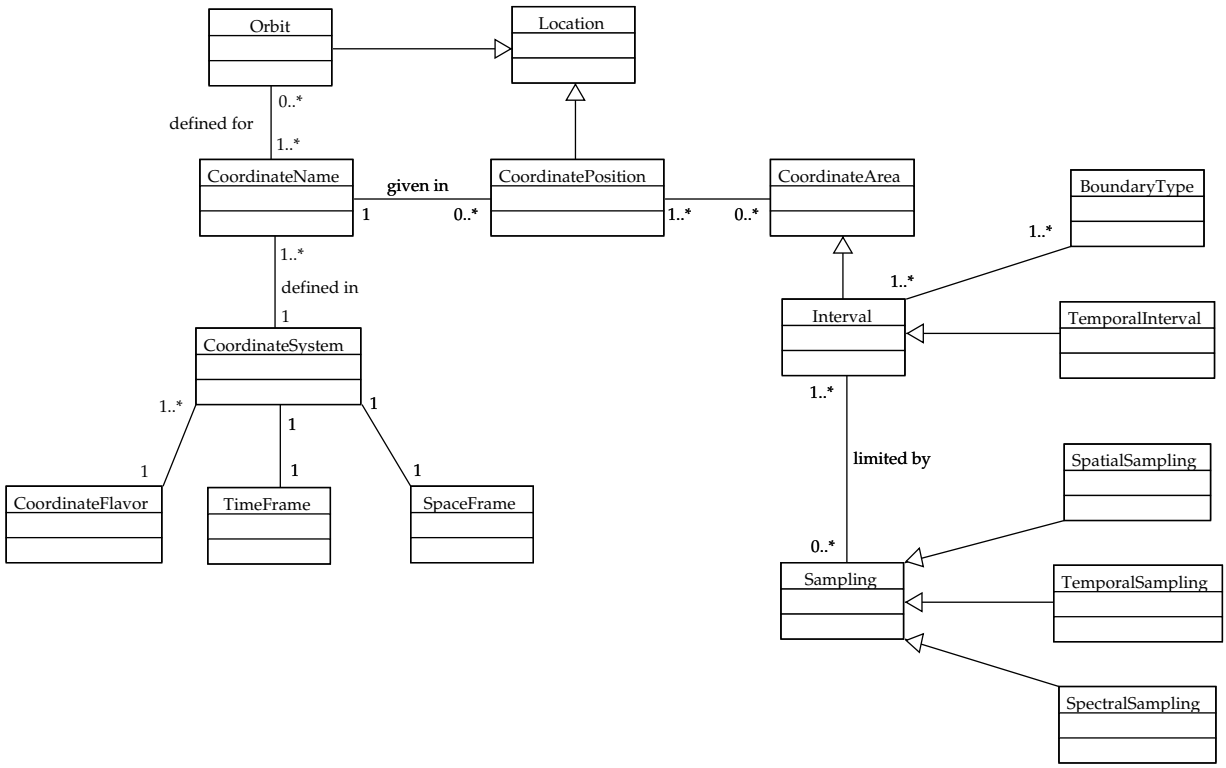


Figure 2: General Metadata Class Model

5.2 Instrument Model

The Instrument Model describes the elements of the data model related to the description of an instrument and its auxiliary components. An instrument is the combination of certain essential elements that allow for the measurement of photons or particles from the Sun or heliosphere in order to provide a sampling in one or more physical observables. An instrument is considered to acquire data in some stable manner, possibly utilizing a number of different configurations. Instruments can be described according to general terms in broad use in the solar and heliospheric domain that help to classify the type of data obtained.

An auxiliary component instead is an element of the overall acquisition system that is not necessarily inextricably bound to an instrument. Such components may be combined with an instrument to construct a working system for data acquisition. Examples of such auxiliary components may include, a telescope, detector, or adaptive optics system. Some physical instruments may be best described with explicit specification of the auxiliary components, while for other, more tightly integrated systems, no specific division into multiple components is necessary. Both instruments and auxiliary components are generally restricted by physical principles to operating in some limited number of electromagnetic energy regimes.

Condensed Description:

An **Instrument**, which can be classified according to a list of general **Instrument Types**, obtains measurements related to one or more **Physical Parameters**. The **Instrument** may provide discreet measurements of the incident photons or particles according to one or more general **Sampling Methods**. The **Instrument** may be utilized in one or more different **Observation Modes**. An Instrument's coverage in a specific **Coordinate** may be described by a **Distribution**. A Distribution may be given as a **DistributionDefinition** that analytically defines a region or a **DistributionMask** that provides a direct map of the coverage. A special case of a Distribution is a spectral **Filter**, which is often given a proper name or described with some specific attributes. An Instrument may optionally be combined with one or more **Auxiliary** components, such as a **Telescope**, to produce a complete **Assembly** for the acquisition of data.

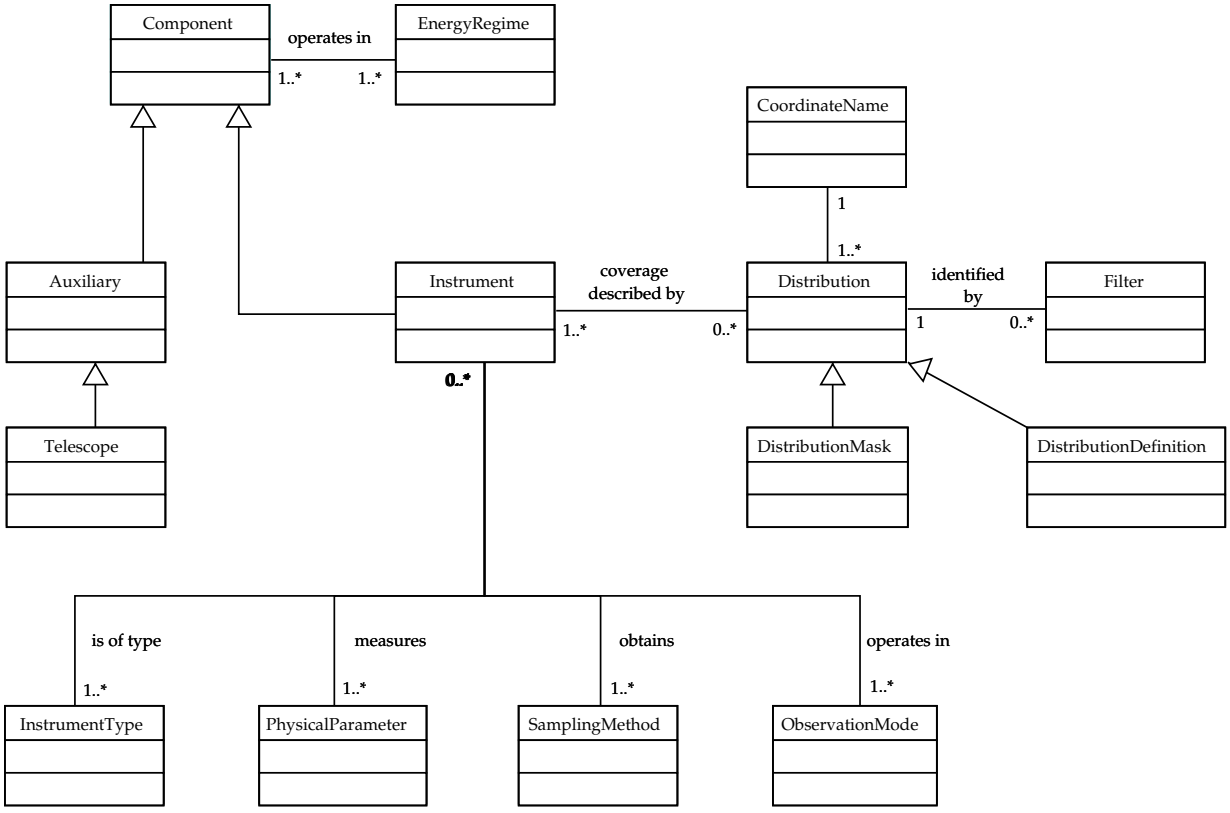


Figure 3: Instrument Class Data Model

5.3 Data Production Model

The Data Production portion of the data model is concerned with the generation of datasets. A dataset, can be produced in several ways: through the recording of primary data; operating on primary data to generate of secondary data products; the modeling of the solar plasma to produce theoretical data. A portion of the Data Production model is concerned with how instruments and other components are joined, both physically and organizationally, under the responsibility of organizations, to form those entities which record datasets. This data model encompasses significant flexibility to permit the description of the range of complex groupings that can be arranged by the data producers. Another portion of the Data Production model describes the processes that can be applied to data.

The model distinguishes between primary and secondary datasets to emphasize the conceptual distinction between the so called raw data and more processed data structures. Primary data are the recorded data in their "rawest usable form", as far up the acquisition chain as it is reasonable to arrive. Processes may have been applied to the data are either essentially free of physical interpretation (e.g. reformatting spacecraft telemetry into proper data files) or are irreversible (e.g. automatic processing that discards the original measurements after extracting the parameters of interest). Secondary data can have an arbitrarily complex amount of processing applied and may combine data from multiple instruments or networks.

Condensed Description:

An **Instrument** may be optionally combined with one or more **Auxiliary** components to form an **Assembly**. For some **TemporalInterval**, this entity must be installed in a **Deployment** at an **Observatory**. It must be possible to determine the **Location** of an Observatory at any given time. One or more Deployments can be combined, for a separate TemporalInterval, into groupings that have some commonality of purpose, called a **Network**. Either a deployment or a network can be defined as the **Recorder** of a **Dataset**. **Organizations** exist which have a possibly shared **Responsibility** for a given TemporalInterval over the Assemblies, Networks, or Observatories.

A **Process** might be applied to a dataset that will modify the content or organization of those data. The underlying component of a Process is the specific installation as a **Package** of a general **Program**. The specific application of the Package may be governed by variable parameters that make up the **ProcessConfiguration**. The composed Process will make alterations to the data content, representation, or storage that can be classified according to a collection of **ProcessEffects**. The Process may also be described by the **ProcessType**, a domain-specific categorization of the actions carried out by the Process.

A possible separate entity, the **Generator**, may apply Processes to one or more Datasets in order to generate additional **SecondaryDatasets**. These secondary products may combine or utilize multiple Datasets. In part, the SecondaryDatasets may be described in the same way as a Dataset.

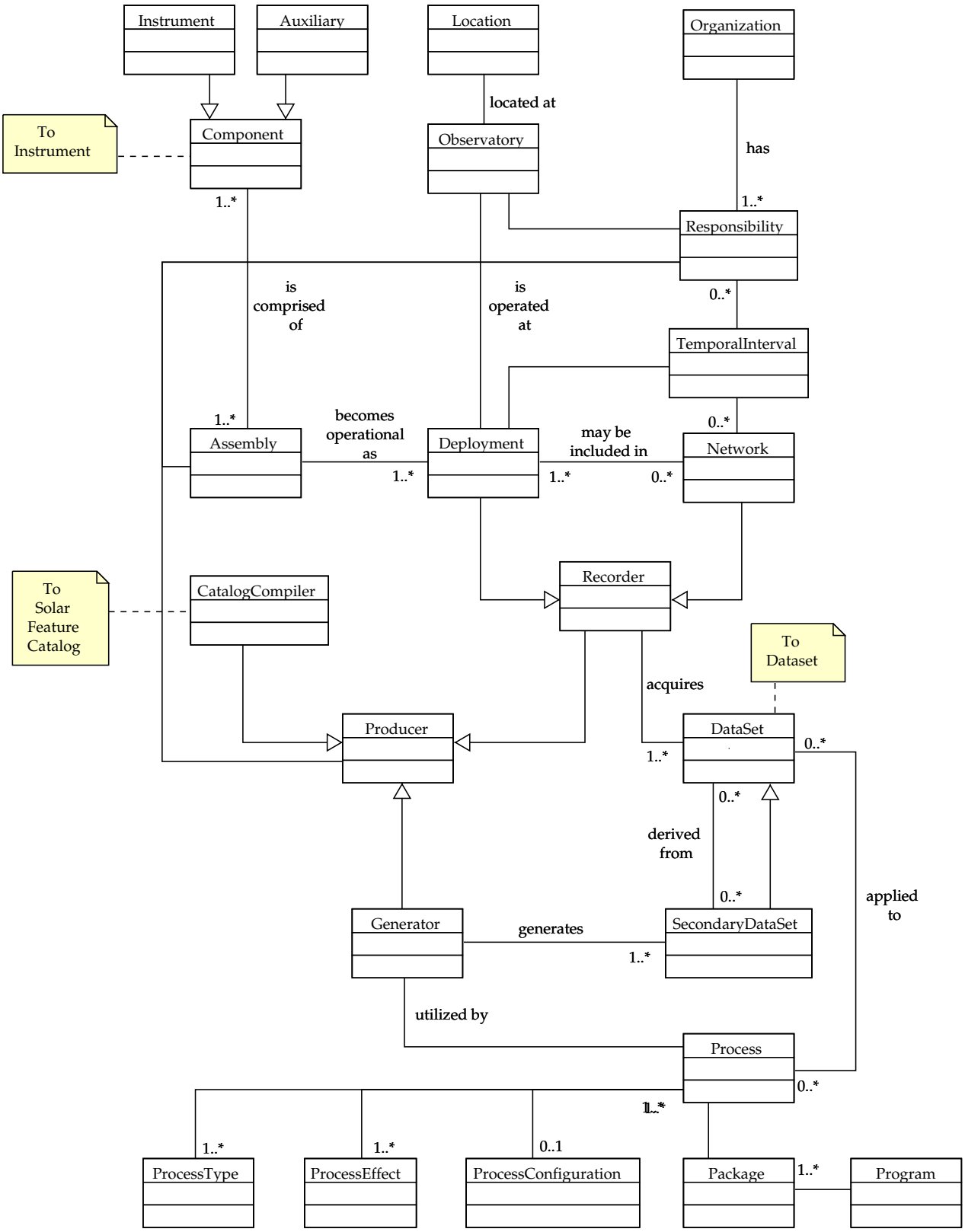


Figure 4: Data Production Class Model

5.4 Dataset Model

The Dataset model describes the metadata concerning the primary data as collected or the secondary data derived from the primary data. These metadata enumerate the coverage obtained by the data, as well other descriptive parameters on which searches might be performed.

A central part of the dataset model is the description of coverage of a dataset in different dimensions. As described in Section 2.6 above, any solar observation can be characterized by the ranges it covers in these essential physical dimensions, time, space, and wavelength. Secondary datasets may present data in other coordinate axes. The definitions of these ranges are the essential descriptive elements of an observation and can be expected to exist for all observations, regardless of the acquisition techniques utilized. Such ranges are among the most common terms employed by the users in the majority of data searches.

In addition to the ranges in the physical dimensions, we can also describe how the sampling of a given dimension was performed. This allows a further characterization of the observation, allowing discrimination among different observations based on these additional statistical quantities. While we concentrate on a parameter space that encompasses these physical spatial, temporal, and spectral dimensions, the technique could be applied to any continuous dimension, such as velocity, density, or frequency, that may be present in secondary data derived from analyzing the raw observations.

Coverage Description for Primary Data

	Temporal	***Spectral***	***Spatial (x_j)***
Start	DATE-OBS	WAVEMIN	X(j)MIN
End	DATE-END	WAVEMAX	X(j)MAX
Count	NSAMPLES	NWAVES	NX(j)POS
Coverage Factor	COV-FAC	WCOV-FAC	X(j)COV-FAC
Regularity	REG-FAC	WREG-FAC	X(j)REG-FAC

Condensed Description:

A **DataSet**, as generated by a data producer, records information that has covers a certain finite volume in one or more coordinate axes. The details of the extent of the covered volume and how the volume is partitioned to provide measurements of different parameters are described by multiple **Samplings**. The coverage in one or more axes may also be described more specifically with a **Distribution**, given either as a region definition or a coverage map. A DataSet may also be described by a **SamplingMethod** and **ObservationMode** which provides domain classifications of the data types. The DataSet description also includes the **Physical Parameters** that are included or derivable from a dataset. Multiple **Quality** descriptors may be given for the dataset, perhaps in terms of the **Resolution** of the data or the **Seeing** conditions during the observations. The ScientificObjective that guided the acquisition of the data may also be described, in particular in terms of the coordinated observation **Campaign** of which the acquisition was part, or the general solar or heliospheric **Object** that was being studied. The identification of a specific structure observed during the observations is given as a **Target**.

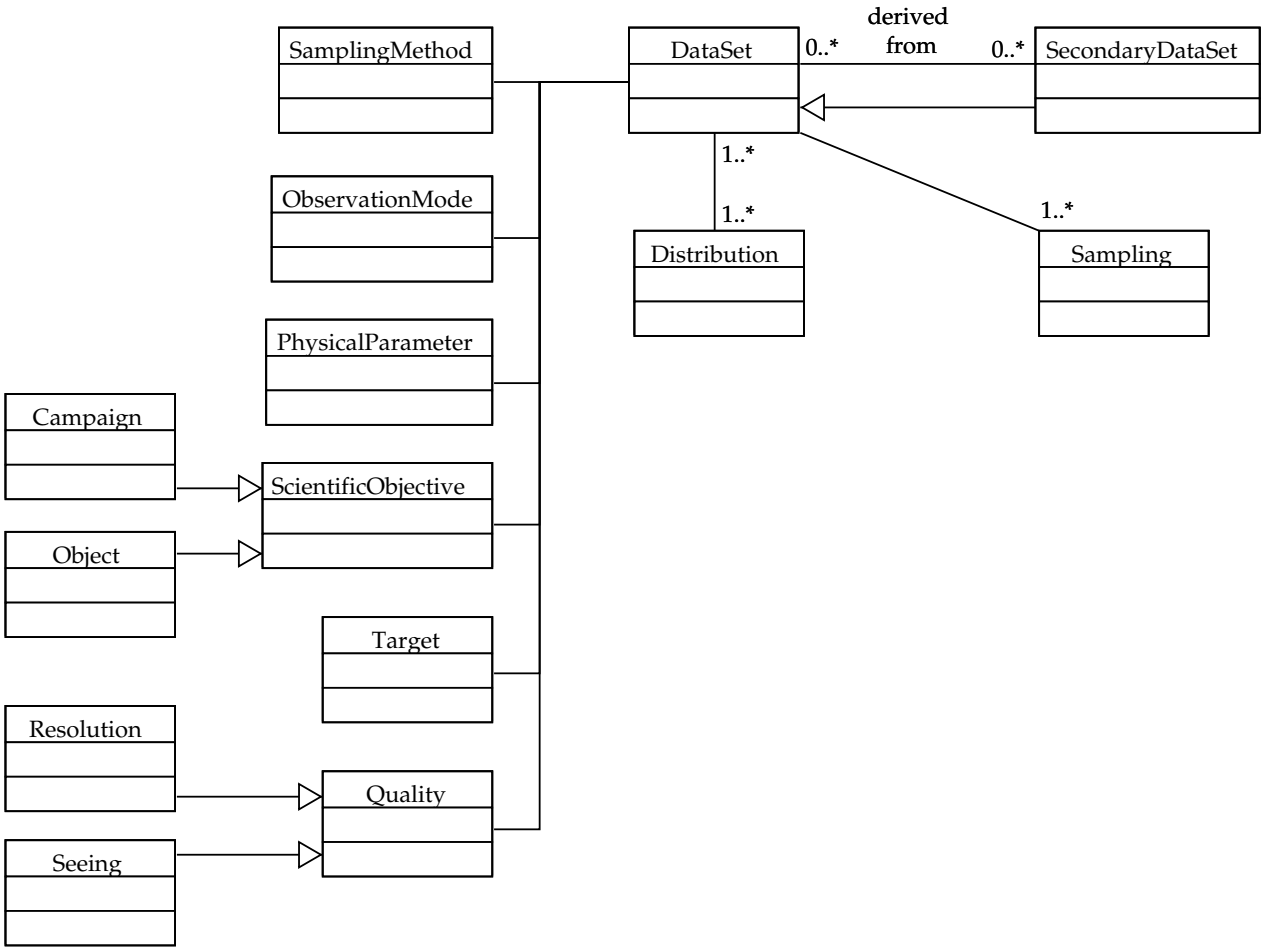


Figure 5: Data Set Class Model

5.5 Data Provision Model

Once a dataset has been recorded or generated, the role of making those data or their associated metadata available on a public or restricted basis belongs to a data provider. The Data Provision model covers the means by which the data are managed and the methods by which they are published. The Data Provision model also covers the publication of catalogs of derived metadata.

Condensed Description:

An **Archive**, a specific type of **Organization**, has the **Responsibility**, for some **TemporalInterval**, of maintaining access to one or more **Providers**. These Providers offer **PublishedMaterial**, which may be a **PublishedDataset** or a **PublishedCatalog**. This material is offered in one or more **DataFormats**, whose generation might be associated with a certain **Process**. Access to the PublishedMaterial may be regulated by the means of **AccessControls**. The accessibility to the stored data may be described by the **DataAvailability.** Both the individual elements of PublishedMaterial and the Provider as a whole may be identified by a Uniform Resource Identifier (**URI**).

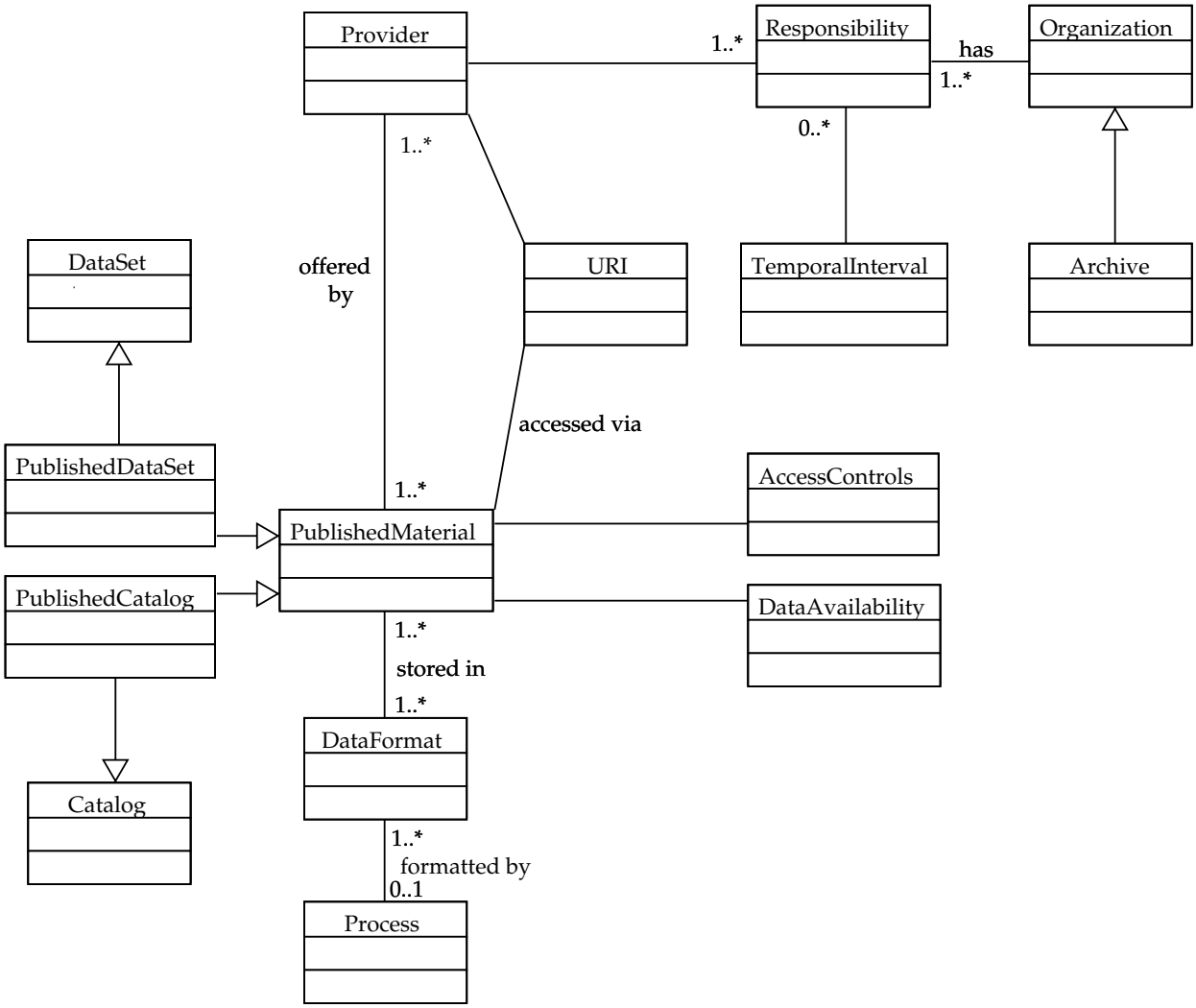


Figure 6: Data Provision Class Model

5.6 Solar Feature Catalog Model

The derived metadata about solar features and events are described according to the Solar Feature Catalog Model. These occurrences, as described in Section 3.4, are observed and information about each one is extracted into tabulations describing some particular grouping of solar features. For each feature, there may be multiple details given, both quantitative measurements and defined classifications. There must be a time associated with each entry since these occurrences are temporally variable.

Condensed Description:

A **CatalogCompiler**, a type of Producer, generates a **Catalog** using one or more **Techniques** to extract the information from the underlying solar data. The Catalog may also be identified with the **Processes** used in its generation. A Catalog will contain information on multiple **Features**, each of which may be identified by a certain **FeatureID**. A recorded feature will be classified as a certain **FeatureType**, which is are specific instances of the general solar and heliospheric **Objects**.

Each Feature may be described by multiple **Entries** related to that feature. Each Entry will be composed of a combination of elements describing the feature's temporal, spatial, and spectral positions, collectively referred to as **EntrySpecifications**. Each Entry must include at minimum at least one **TimeEntry**, describing the moment for which a given

entry applies, specified by the **TimeEntryType**. The Entry may also contain one or more **PositionEntries** and **SpectralEntries** defining the location and spectral range in which the feature was observed. All these positions, that may define a boundary of the **EntryBoundaryType** for the observed feature, can be described by a simple **CoordinatePosition**, a composed **CoordinateArea**, or an exact coverage given by a **Distribution**. For each entry, it is also possible to associate a **Measurement**, for measured quantities, or a **Classification**, for non-quantitative descriptive terms. A simple Boolean flag for a given quality is given as an **EntryFlag**. The means by which Measurements or Classifications of specific type are sorted is given by an **Ordering**. Each Entry may be associated with the one or more Datasets from which the Entry information was derived.

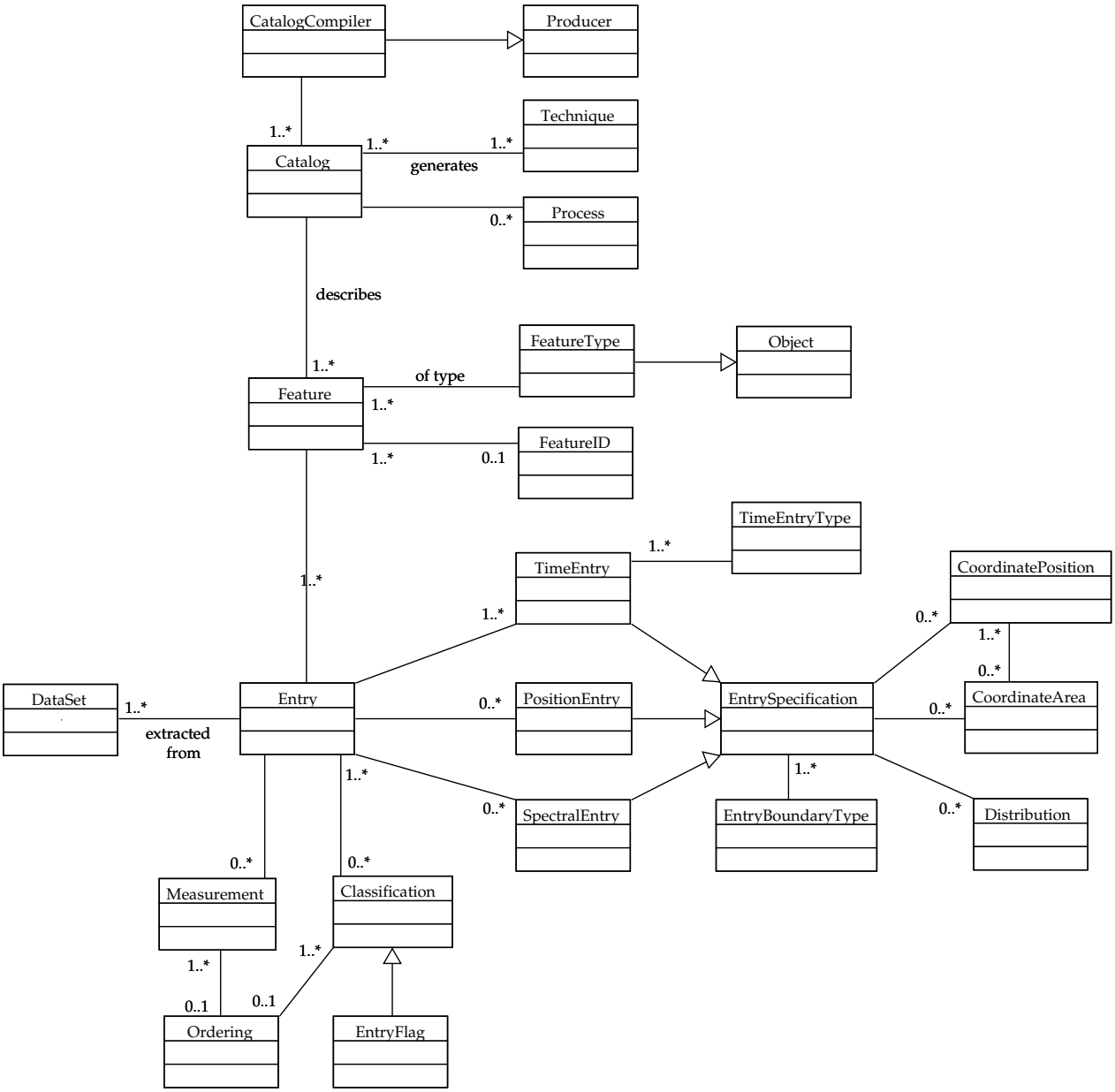


Figure 7: Solar Feature Catalog Class Model

6. EGSO Data Model Implementation

In order for EGSO to function, it must handle a wide variety of metadata from a variety of sources and a range of formats. EGSO will use the data model to guide the interpretation of these heterogeneous metadata. Other metadata may be mapped into *registries* that would contain information concerning specific players or material available to the system. The data model will provide a common underlying structure on which to base different metadata components.

This data model will be employed at various points throughout the system. The identification of resources of interest might use the elements of the model describing data producers to perform the desired search (e.g. find only UV instruments operating from during the past 10 years). The automated application of analysis or image mapping software would use the data model to determine the validity of the attempted operation. And of course, the searches for datasets of interest, the core of EGSO, would operate on those elements of the data model tailored to the description of solar datasets, the Unified Observation Catalog.

6.1 Unified Observation Catalog

In order to enable the tools that will allow EGSO to provide truly new capabilities to solar physicists, facilitating innovative uses of the available data, it is necessary to provide a common description with which to bring together the diversity of observations. In order to achieve the goals of the project, it would not be sufficient to merely provide access to data resources and transmit the contents of those resources, unaltered, directly to the end user. This would not provide for the joint searches across multiple catalogs or a more uniform presentation of the returned results. Therefore, the content of the observation catalogs from the data providers must be converted and translated into some common framework.

6.1.1 The UOC Concept

This framework is the Unified Observation Catalog (UOC), which will function as a single catalog containing descriptions of all incorporated solar observations. The UOC will not be a replacement for the individual catalogs maintained by each instrument team, but rather would provide a layer of abstraction above these archives in such a way that all existing archives could be searched, joined, and perhaps presented in a common format. This common representation would facilitate the searching and utilization of multiple datasets by the scientist and reduce the barriers to including additional data sources in their work.

Because the UOC is conceived as a layer above the existing catalogs, it will reflect the quantity and quality of the information in the existing catalogs. The amount of detail and level of granularity varies among catalogs from different producers, some listing detailed information for individual exposures, others simply providing condensed information about extended sequences or entire observing days. Rather than reduce the information from all catalogs to a "lowest common denominator" level of detail, which would risk discarding too much useful description of some datasets, the UOC should be able to incorporate catalogs with heterogeneous levels of detail. In its most detailed realization, the UOC should contain information at the finest level of granularity as provided by each catalog.

6.1.2 Catalog Mapping onto the UOC

Conceptually, the process of generating the UOC will involve an initial description of a given catalog's content and organization, followed by the automatic interrogation of the catalog guided by this general description. The first step presumably requires human intervention to identify the correspondence between the particular labels utilized by a specific catalog and the general terms that are part of the common description in the UOC. This *mapping* will be codified and stored in a common format (e.g. an XML schema) for repeated use within the system. The catalog description will allow for the automated formulation of queries and interpretation of the returned results from each catalog.

In this process, any incomplete or incorrect metadata present in the underlying catalogs will be carried forward, not eliminated or hidden. Some deficiencies may be corrected as part of the analysis to produce the mapping, but information that is simply not present, or contains unpredictable errors, cannot be remedied easily. It will essentially be the responsibility of the individual providers to generate a valid catalog for inclusion into the system. It is probable that the additional scrutiny placed on the catalogs in order to incorporate them into this broader scope will highlight problems that had heretofore not been of significance. The value of having those catalogs merged into the unified catalog together with a wealth of other information may provide the motivation to make the needed corrections.

6.1.3 Information Content of the UOC

It should also be noted that the UOC will not necessarily include all the information contained in the underlying catalogs. Some catalogs contain information on instrument specific parameters that, while valuable in the context of that instrument, are not of general interest in performing multiple resource data searches. Since the UOC will not replace the instrument specific archive, this information will still be available directly from the instrument catalog itself externally from the UOC. Instrument specific parameters which have a known impact on the data themselves, for example an instrument setting that alters resolution or data quality, could be "translated" for inclusion in the UOC in terms that more generally reflect the effects produced on the data. This would make such information more widely useful by eliminating the instrument specific terminology.

In keeping with the goals of EGSO, the UOC is also intended primarily as a catalog for performing data searches, rather than being designed for reduction of the associated data. The information contained in the UOC *may* be sufficient to perform certain actions on the data, but it will not be a requirement that the UOC contain all the information necessary to manipulate a given dataset. It may be left to an instrument's particular software package to reconstruct or perform operations on a given dataset obtained with that instrument (e.g. generate an image from a rastering instrument for overlay with other datasets).

The UOC is a construct that is aimed to facilitate the operation of EGSO. The data are described in a manner that will facilitate search operations and intercomparisons among multiple heterogeneous datasets. It is not directly intended as the representation that will be presented to all end-users. In the generation of the user interfaces to the system, it should be possible to operate on the information stored in the internal representation of the UOC to produce parameters best suited to a particular application.

The UOC is a general purpose catalog that must attempt to maintain valid representations of a wide diversity of multiple types of data from numerous instruments. This will require, to an extent, an abstracted description of the contents of the underlying observing catalog. Given that there is no primary type of data to be accessed, no one standard for catalog format or content, no general understanding of the definition of an observation, such an abstraction will necessarily bear apparent differences from all existing catalogs. The goal in constructing the UOC is not to maintain fidelity to a particular catalog format over another, but to minimize the amount of information that might be lost in this adaptation.

The Unified Observation Catalog as discussed here has concentrated on the description of *primary data*, the actual observations from the Sun and heliosphere. However, there are many *secondary data* that come from the processing or combination of primary data. These data products will also play an important role in EGSO and not all work requires accessing the primary data itself. Because these secondary data are directly derived from the primary data, there should be much overlap in their descriptions that will allow the UOC to incorporate these data as well.

6.2 Solar Feature Lists

In addition to the observation catalogs, there are catalogs which attempt to compile information extracted from the observations in order describe actual occurrences in the solar atmosphere. The metadata derived from the contents of the primary or secondary data are called *derived metadata*. These catalogs can be generated both through manual examination of the data as well as through automatic processing utilizing various techniques.

Analogously to the situation with the varied representations found among observation catalogs, the derived metadata catalogs are produced in a variety of formats with vastly different levels of detail in their content. Because of this heterogeneity, it is often difficult to make intercomparisons of the content multiple catalogs, or to combine the content of these catalogs with the observation catalogs beyond the simplest of searches. Here too, a data model is needed that will allow these occurrences to be described in some common framework. This framework must be consistent with the rest of the data model, in particular the UOC, so that the contents of the derived and observation catalogs can be compared.

There are two primary classes of features that may be identified in the solar atmosphere or heliosphere. There are *events*, which are rapidly evolving phenomena that are primarily defined by their temporal lifetime or evolution. There are also *features*, which are more stable phenomena (yet these too will evolve, albeit on longer time scales), that may be observed on multiple occasions during their lifetime.

EGSO will have to handle both event and feature catalogs in the application. In addition, it will face simple catalogs generated with extensive human intervention, as well as richer catalogs with information generated by automated feature recognition techniques. The model for a common description of these derived metadata catalogs will have to able to cope with the information arriving from numerous sources.

7. Data Model Realizations

The data model described above is an abstract description of the organization of the metadata that are to be included in the model. However, there will be multiple realizations of this data model for uses in different components of EGSO. Some of these realizations will include:

- FITS Keywords stored in data files;
- Data Catalogs stored in relational or other databases;
- Catalog fragments exported to XML files;
- User interface elements presenting the metadata to the user;
- User-generated catalogs of derived metadata;

7.1 FITS Keywords for Data Files

One area where increased attention to interoperability could yield great gains, both within the scope of EGSO as well as in solar physics as a whole, would be the usage of FITS keywords in a more coordinated manner. Since the FITS keywords are a primary means by which certain observational metadata are recorded, the adoption of a common definition for some general concepts would automatically lead the way to more unified description of those metadata. Therefore, from the data model described above, we propose the following concepts and terminology as a FITS keyword convention for use within solar physics. All the concepts defined in the data model will not necessarily be found in the list of FITS keywords below, since certain concepts (often those that are more static) are not generally recorded in the FITS headers. The use of such a convention should never be seen as mandatory for inclusion in EGSO, but rather is an optional measure that would provide the data producer with a wider compatibility with various catalogs and software tools.

As noted in Section 4, there has already been some progress in this area. In particular, the definition of the FITS keywords used by SOHO [3] has often been used outside of the initial scope of the SOHO instruments. There are certain concepts which are not treated in the SOHO list and some keyword definitions have not seen widespread adoption. Nonetheless, it is from this list, already widely supported in the community, that we draw some common definitions.

There are two aspects of the usage of FITS keywords that produce problems and ambiguities in the interpretation of the associated metadata. The first is the usage of different keywords for expressing the same concept. In this case, two FITS files may both incorporate the same information but under different labels (e.g. EXPTIME, and TEXP). Though it is possible to provide translation services that have knowledge of the synonyms among datasets, it would be preferable to use a common keyword among multiple projects to describe the same concept. A more difficult problem is that in which the same label is used to describe two different concepts within the metadata realm (e.g. OBS_MODE that may describe an instrumental configuration or motivational aspects of the data acquisition). This problem requires more attention in the translation process to resolve this overlap and generate a valid set of metadata. Obviously, it would be preferable if both of these problems could be avoided.

7.1.1 Coordinate Definitions

One area in which it is very valuable to have common metadata descriptions is the definition of the instrument field of view or location in some specified coordinate system. There are two common forms for representing coordinate systems in a FITS header. The first method is uses the keywords CPRIX, CTYPE, CUNIT, etc. as defined in the original FITS standard [4] for arrays of data in one or more coordinate axes. Recently, a more complete model for describing the coordinates of a data array and possible transforms to alternate coordinate systems has been defined as the World Coordinates System [5][6]. The WCS formalism allows more complex transformations to be described and the definition of multiple coordinate systems for any given dataset. It is expected that both forms will continue to be used in the future in solar physics, depending on the interest of the data producer and the perceived needs for a particular instrument.

Thompson [2] has described different coordinate systems and their representation in FITS headers. For many observations, heliographic or Carrington coordinates are sufficient and are especially suitable for catalogs of solar features. Heliocentric coordinates may be preferred for some observations given their natural relation to the detector coordinates and extensibility to off-limb observations. Helioprojective coordinates may also be suitable for some instruments. Since it is possible to convert among any of these coordinate systems if they are properly specified, any of these coordinate systems are equally valid. It is essential, however, that all the necessary FITS keywords, as specified by Thompson, are included in the header to properly define the pointing.

7.1.2 Time Systems

There are different time scales that be used to temporally locate an observation. The most commonly used time system is Coordinated Universal Time (UTC) and will presumably remain the principle time system of reference in solar physics. In FITS headers, the time scale is assumed to be in UTC unless otherwise specified. Alternative time systems can be specified using TIMESYS keyword to identify the time scale. Alternative time scales that might be used in some special cases include International Atomic Time (TAI), which is a time scale without leap seconds, and Terrestrial Time (TT), the IAU standard time scale and has a fixed offset from TAI. It should also be noted that the definition of UTC may be changed in the future (elimination of the leap seconds), and a new time scale may be defined.

Instrument Related:

Data Model Class	Related FITS Keyword	Keyword Definition
Instrument	INSTRUME	The value field shall contain a character string identifying the instrument used to acquire the data associated with the header. (i.e. “CDS”, “SUMER”, etc., for SOHO). The instrument is the combination of essential elements to allow for the measurements of photons or particles.
Auxiliary	TELESCOP	The value field shall contain a character string identifying the telescope used to acquire the data associated with the header (i.e. SOHO). The telescope is the support where instruments can be mounted.
	DETECTOR	Detector used to acquire the data.
EnergyRegime	OBS_TYPE	The spectral domain in which the observations were made.

Data Production Related:

Data Model Class	Related FITS Keyword	Keyword Definition
Observatory	OBSERVAT	A character string identifying the physical entity, located in a defined location, which provides the resources necessary for the installation of an instrument.
		.
Organization	INSTITUT	Name of an institution or organization that plays a role in acquiring a dataset.
Network	NETWORK	Organizational entity of a series of instruments with similar characteristics or goals that operates in some coordinated manner or produces data with some common purpose.

Data Provision Related:

Data Model Class	Related FITS Keyword	Keyword Definition
PublishedData	FILENAME	The name of the data file
URI	URL	A URL giving the location from which the data can be accessed
Organization	INSTITUT	Name of an instrument team or institution taking part in a coordinated observing program

Data Set Related:

Data Model Class	Related FITS Keyword	Keyword Definition
Sampling *General*	NAXIS*n*	The value field shall contain an integer ranging from 1 to 999, representing one more than the number of axes in each data array.
	COUNT	Number of repeated observing sequences within an observing program.
Sampling *temporal*	DATE–OBS	Start date of observing program (UTC).
	DATE–END	End date of observing program (UTC).
	EXPTIME	Duration of exposure, in seconds, where exposure is considered the effective amount of time for which the detector was measuring photons or particles coming from the source of interest.
	TIMESTEP	Estimated temporal step between two exposures, in seconds.
	CADENCE	If observations are repeated cyclically, the approximate number of frames acquired per hour.
Sampling *spectral*	WAVEMIN	Minimum wavelength of observation, in *nm*.
	WAVEMAX	Maximum wavelength of observation, in *nm*.
	WAVEUNIT	The units of the wavelength.
	WAVELNTH	The wavelength of observation, in *nm*. When used for observations that cover a range of wavelengths, then this keyword represents the wavelength of interest, not necessarily the central wavelength
Sampling *spatial*	CENTER_X	The coordinate of the center of the Sun in pixels along the first dimension, where the center of the first pixel in the image has the coordinate value 1 along each axis
	CENTER_Y	The coordinate of the center of the Sun in pixels along the second dimension, where the center of the first pixel in the image has the coordinate value 1 along each axis
	IXWIDTH	Maximum width of the instrument field-of-view in the instrument X axis, i.e. the direction perpendicular to the vertical axis as used in keyword ANGLE
	IYWIDTH	Maximum width of the instrument field-of-view in the instrument Y axis, i.e. the direction along the vertical axis as used in keyword ANGLE
	XCEN	Center of the instrument field-of-view along the solar X-axis
	YCEN	Center of the instrument field-of-view along the solar Y-axis

	ANGLE	Angle of rotation of the vertical axis of the instrument field-of-view relative to solar north
Sampling *Global acoustic*	SHDLMIN	Minimum value of spherical harmonic degree ℓ
	SHDLMAX	Maximum value of spherical harmonic degree ℓ
	SHDLSTEP	Spacing between spherical harmonic degree ℓ
	SHONMIN	Minimum value of spherical harmonic radial order n
	SHONMAX	Maximum value of spherical harmonic radial order n
	SHONSTEP	Spacing between spherical harmonic radial order n
Sampling *Local Acoustic*	K_UNIT	The units of the spatial wavenumber
	KX_MIN	Minimum value of the x-component of the spatial wavenumber
	KX_MAX	Maximum value of the x-component of the spatial wavenumber
	KX_STEP	Sampling size of the x-component of the spatial wavenumber
	KY_MIN	Minimum value of the y-component of the spatial wavenumber
	KY_MAX	Maximum value of the y-component of the spatial wavenumber
	KY_STEP	Sampling size of the y-component of the spatial wavenumber
	KZ_MIN	Minimum value of the z-component of the spatial wavenumber
	KZ_MAX	Maximum value of the z-component of the spatial wavenumber
	KZ_STEP	Sampling size of the z-component of the spatial wavenumber
Sampling *Temporal frequency*	FRQUNIT	The units of the temporal frequency
	FRQMIN	Minimum value of the temporal frequency
	FRQMAX	Maximum value of the temporal frequency
	FRQSTEP	Sampling size of the temporal frequency
Coordinate	CRPIX*n*	The value field shall contain a floating point number, identifying the location of a reference point along axis n, in units of the axis index. This value is based upon a counter that runs from 1 to NAXISn with an increment of 1 per pixel. The reference point value need not be that for the center of a pixel nor lie within the actual data array. Use comments to indicate the location of the index point relative to the pixel.
	CRVAL*n*	The value field shall contain a floating point number, giving the value of the coordinate specified by the CTYPEn keyword at the reference point CRPIXn.

	CDELT*n*	The value field shall contain a floating point number giving the partial derivative of the coordinate specified by the CTYPEn keywords with respect to the pixel index, evaluated at the reference point CRPIXn, in units of the coordinate specified by the CTYPEn keyword.
	CROTA*n*	This keyword is used to indicate a rotation from a standard coordinate system described by the CTYPEn to a different coordinate system in which the values in the array are actually expressed. Rules for such rotations are not further specified in this standard; the rotation should be explained in comments. The value field shall contain a floating point number giving the rotation angle in degrees between axis n and the direction implied by the coordinate system defined by CTYPEn.
	CTYPE*n*	
	CUNIT*n*	
Target	OBJ_ID	Object identifier, e.g. active region number
	TARGET_ID	The physical entity that is the focus of a particular observation
ObservationMode	OBS_MODE	Observing mode
	OBS_TYPE	The spectral domain in which the observations were made
Quality	QUALITY	The percentage of data obtained that is "good" (e.g. not affected by the South Atlantic Anomaly for spacecraft or by clouds for ground-based observatories)
	SEEING	A measurement of the seeing conditions at the time the data were acquired
	SEQVALID	Either “Y” or “N”, denoting whether or not the data from an observing sequence has scientific value
ScientificObjective	OBJECT	Character string containing the name of the object observed, e.g. “CORONAL HOLE”, etc.
	CAMPAIGN	Description of the observation campaign or Joint Observing Program (JOP) under which the observations were made
Distribution	FILTER	Filter used to acquire the data.
	FIELVIEW	The dimensions of the field of view of the image, in arcseconds *x* arcsecond
PhysicalParameter	PHYSPARA	Physical parameter represented in the data array.

References

1. Booch, G., Rumbaugh, J., Jacobson, I.: 2000, *Introduction to the Unified Modeling Language*, Addison Wesley Longman.

2. Thompson, W., 2001, *Coordinate Systems for Solar Image Data*.

3. Howard, R., Thompson, W.: 1995, *SOHO FITS Keywords*

4. Wells, D.C., Greisen, E.W., Harten, R.H.: 1981, *Astron. Astrophys. Supp.*, **44**, 363.

5. Greisen, E. W., Calabretta, M. R.,: 2002, *Astron. & Astrophys.*, **395**, 1061.

6. Calabretta, M. R., Greisen, E. W.; 2002, *Astron. & Astrophys.*, **395**, 1077.

Glossary of Terms

Content Metadata – descriptive elements that provide information about the data to which they are related.

Curation Metadata – descriptive elements detailing how the related data are stored or accessed.

Data Producer – an entity that acquires or generates primary data.

Data Provider – an entity that makes data available in some format through an electronic interface.

Derived Metadata – metadata extracted from Primary Data through subsequent analysis or processing, e.g. the SFC and SEC

Metadata – any Data that provides information about Data (including other metadata).

Primary Data – archived information that represents the baseline storage of an observation or measurement. These are the data in their "rawest" form, which may be raw photon counts, or more often "data numbers" which is the result of converting the incoming flux to a digital representation.

Secondary Data – any subsequent representation of the data that is the result of processing (summarizing, modifying, transforming, etc.) Primary Data. A datum may still be considered as Secondary Data even if the corresponding Primary Data is not associated with the system.

Appendix A: Common Observational Modes

A1.1 Classical Imaging

Many instruments operate by creating an image of the Sun in two spatial directions as projected on the plane of the sky from the observing location. The image may cover the entire solar disk (i.e. *full-disk*), a limited region on the solar disk, or portions of the corona extending beyond the visible disk (or some combination of these). The spectral information is limited to wavelength bands of differing widths and shapes as defined by filters or other elements that affect the instrument response. These images are generally obtained with two-dimensional detectors, primarily charge-couple devices (CCD). The temporal coverage for a single image is defined by the beginning and end of the exposure during which the detector was exposed to the incoming photons.

Dimension	Coverage	Definition
Spectral	*Range*	Minimum and maximum wavelengths: λ_{min}, λ_{max} ($\Delta\lambda = \lambda_{max}-\lambda_{min}$) Central wavelength ($\lambda_0$) $\approx \Delta\lambda/2$
	Number of Elements	1 wavelength band
Temporal	*Range*	Start and end exposure: t_{min}, t_{max} ($\Delta t = t_{max} - t_{min}$)
	Number of Elements	1 individual exposure (Δt)
Spatial	*Range*	*Rectangular:* x_{min}, x_{max};y_{min}, y_{max} ($\Delta x = x_{max} - x_{min}$, $\Delta y = y_{max} - y_{min}$) *Annular:* r_{min}, r_{max};θ_{min}, θ_{max}
	Number of Elements	Number of spatial pixels: N_x, N_y
		Where: [*x, y*] Cartesian coordinates [r,θ] Polar coordinates (Sun-centered)

Classic image array description

A1.2 Classical Spectrography

The radiation coming from the Sun may be spread out by dispersive elements to allow the separation of photons of different energies. In this way, detectors that are not sufficiently sensitive to photon energy can be used to measure the incoming flux at different wavelengths. Often a slit is used to sample a number of spatial points in a direction orthogonal to that of the dispersion, in this way creating a two-dimensional array in λ, x (or y) which can be sampled by a two-dimensional detector. The spectral coverage is given by the dispersion and detector size, while the spectral resolution is determined by the width of the slit, the properties of the dispersive element, and the detector sampling. The temporal coverage is again defined by the duration of the exposure.

Dimension	Coverage	Definition
Spectral	*Range*	Multiple bands: λ_i $i = (1,\ldots, n)$ Minimum and maximum wavelengths per band: $\lambda_{i,min}$,$\lambda_{i,max}$
	Number of elements	$N_{\lambda} = \prod_{1}^{n} (N_{\lambda})_i$ number of spectral pixels
Temporal	*Range*	Start and end exposure: t_{min}, t_{max}
	Number of elements	1 individual exposure
Spatial	*Range*	Multiple spatial bands: x_j $j = (1,\ldots, m)$ Minimum and maximum coordinate per spatial band: $x_{j,min}$, $x_{j,max}$
	Number of elements	$N_x = \prod_{1}^{m} (N_x)_j$ number of spatial pixels
		Where: n – number of spectral bands m – number of spatial bands

Classic spectrographic array description

A1.3 Rastering

An instrument with limited coverage in one of the primary coordinates (e.g. x) may make repeated observations, slightly shifting its position along that coordinate in order to gradually build up coverage in an additional dimension. For example, a classic spectrograph may take multiple measurements, shifting its slit to other spatial positions, generally in a direction perpendicular to the slit orientation. In this way, a two-dimensional map in the spatial dimensions can be obtained, with the spectral information at each point providing a third dimension. Similarly, a tunable filter mechanism can obtain images covering two-dimensions at a series of closely spaced wavelengths, the resultant spectral scan allowing the measurement of spectral information from the resultant 3-D data cube. The acquisition of a time series of repeated images could be considered a raster in the temporal dimension.

Many rastering schemes, including those described above, actually combine the temporal dimension with the other coordinates over which the raster is being performed, though this need not always be the case (e.g. a multi-slit spectrograph).

Dimension	Coverage	Definition
Spectral	*Range*	Multiple bands: λ_i $i = (1,\ldots, n)$ Minimum and maximum wavelengths per band: $\lambda_{i,min}$, $\lambda_{i,max}$
	Number of elements	$N_\lambda = \prod_{1}^{n} (N_\lambda)_i$ number of spectral pixels
Temporal	*Range*	Start and end raster: $\Delta T = T_{max} - T_{min} = t_{p,max} - t_{1,min}$
	Number of elements	1 *or* $N_t = \prod_{1}^{p} (N_t)_k$ $k = (1,\ldots, p)$
Spatial	*Range*	Multiple bands: x_j $i = (1,\ldots, m)$ Minimum and maximum coordinate per band: $x_{j,min}, x_{j,max}$ Multiple scan positions: y_k *or* $y \in [y_1, y_2 \ldots y_p]$
	Number of elements	$N_x = \prod_{1}^{m} \prod_{1}^{p} (N_x)_{j,k}$ $N_y = p$
		where: n – number of spectral bands m – number of spatial bands p – number of raster positions

Scanning spectrograph raster description

Appendix B: Data Model Classes

AccessControl

The general description of the limits imposed by the controlling Archive for access to the associated Datasets. Such controls may request certain information before allowing access (without necessarily verifying that information) or may provide effective restrictions on access to the data.

Appears in: *Resource Registry*

AccessControl
+ name: String + controlType: String

Archive

The structure responsible for holding a catalog or data files for external access. An archive is generally a single entity, managed at one Organization, which may maintain portions of one or more resources.

Appears in: *Resource Registry*

Archive
+ name: String + accessQuality: String + available: boolean

Assembly

An assembly is a combination of components, of both Instrument and Auxiliary types, that form a unique entity for the acquisition of data with certain characteristics. An Assembly must contain one Instrument, and may contain one or more Auxiliary components. The modifications that are made to a data acquisition system (often called *instrument upgrades*) can in some cases be modeled as the generation of a new Assembly object combining the same Instrument object with new Auxiliary objects. In this way, such changes can be properly recorded while maintaining the connection among what are often considered by solar physicists as different versions of a "single" instrument.

Appears in: *Creator Registry*

Assembly
+ name: string

Auxiliary

Auxiliary components are those additional elements that may be used together with an Instrument in the acquisition of data, but are not themselves integral in defining the general means by which the data is sampled (c.f. SamplingMethod). These elements can be logically (and perhaps physically) separated from the essential components of the Instrument. The distinction between component types may also be made to highlight administrative or operational differences among the elements (e.g. different responsibilities among elements). Some examples of possible auxiliary components include telescopes, detectors, and adaptive optics systems.

The decision as to what components to classify as auxiliary components should be done on an instrument by instrument basis. Ground-based instruments, which may undergo frequent modifications, changing detectors or moving between telescopes, might be best modeled as multiple components separating these elements. A space-based instrument, where such changes are less common, may instead be better defined as a single element assembly (except where the description as multiple components might provide additional useful information).

Appears in: *Creator Registry*

Auxiliary
+ name: String + component Type: String

BoundaryType

The limits of an interval may have different meanings, depending on exactly how that limit was determined. Some intervals may be exactly defined, while other intervals are less rigidly constrained due to observational or operational limitations. For example, the determination of the temporal limits for an observed solar feature may not be exact due to limits in the temporal sampling, the day-night observation cycle, or the visibility of the feature due to solar rotation.

Appears in: *Solar Feature Catalog*

BoundaryType
+ name: String

Campaign

An observation may be part of a coordinated observation program whose goal is often to obtain data from multiple instruments in order to obtain additional coverage in the spectral or temporal dimension. These are temporary coordinations of effort have value in producing multiple datasets all obtained with a common purpose, hence increasing the value of the combination of the data from different instruments. These campaigns are almost always organized on an *ad hoc* basis relying on best effort contributions from multiple organizations. There are a variety of different groups that organize such campaigns, the primary ones currently being SOHO with its Joint Operating Programs (JOP) and the Max Millenium group flare campaigns.

Appears in: *UOC*

Campaign
+ name: String + campaignType: String + campaignPeriod: TemporalInterval

Catalog

The Catalog in this context refers to the generated lists of solar events or features that are used within solar physics to help locate data of interest. The Catalog may be produced in real time, or may be generated much later in a more static manner. Each catalog will include information on one or more types solar features generated using one or more analysis techniques.

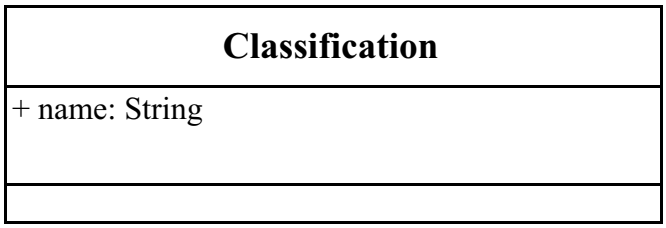
Appears in: *Solar Feature Catalogs*

Catalog
+ name: String + generationInterval: TemporalInterval + catalogUpdateFrequency: String + reliability: String

Classification

A Classification is a qualitative ranking or identification given to an observed feature as recorded in a solar feature catalog.

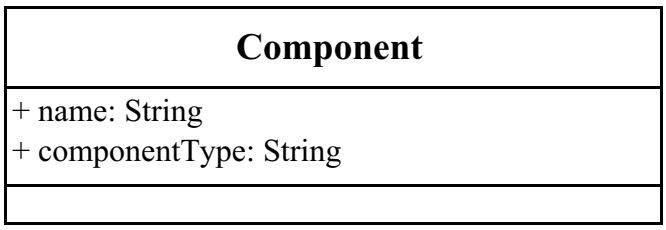
Appears in: *Solar Feature Catalog*



Component

The Component is the superclass that provides a common interface for any entity that can be combined into an Assembly. A Component operates in a limited number of EnergyRegimes.

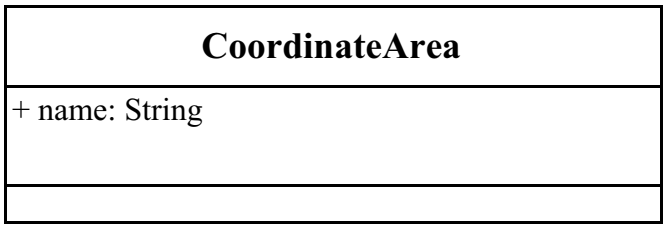
Appears in: *Creator Registry*



CoordinateArea

A general description of regions extending over one or more coordinate axes. A CoordinateArea will be described by one or more CoordinatePositions, together with some area descriptor that indicates how to assemble those points into the desired region (e.g. two positions defining the two opposite corners of a box).

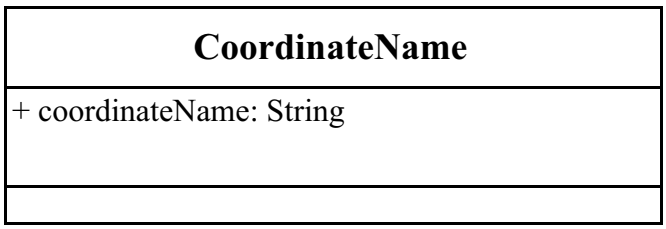
Appears in: *UOC, Solar Feature Catalog*



CoordinateName

Defines the specific axis or axes in a coordinate system in which a particular sampling or distribution refers. For example, may specify the latitude or longitude of a heliographic coordinate.

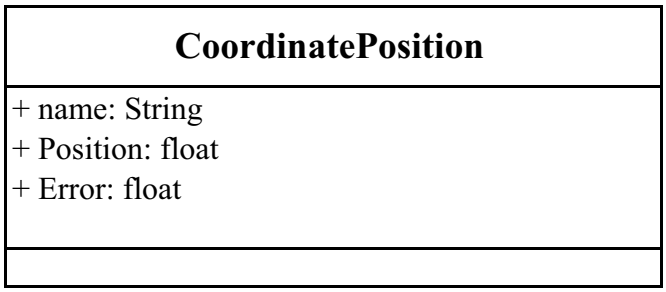
Appears in: *UOC, Solar Feature Catalogs*



CoordinatePosition

A location as defined in some chosen coordinate system. The position may be defined in two, three, or more dimensions. The location need not be only a spatial dimensions, but may be given as well for some other coordinate. The CoordinatePosition may also include information on the errors associated with any position measurement.

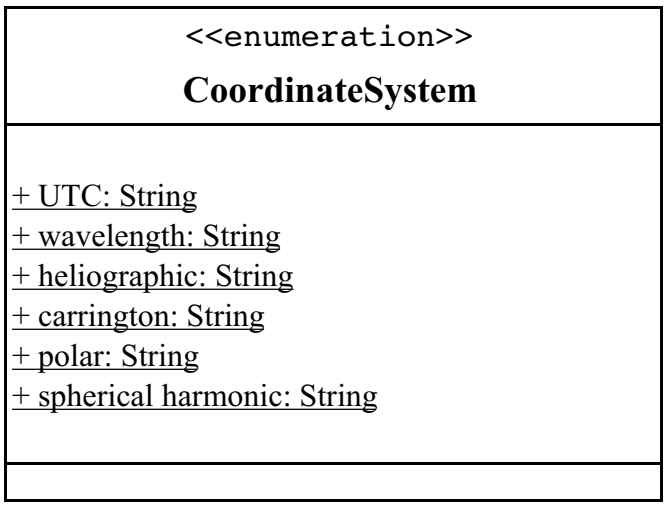
Appears in: *Resource Registry*



CoordinateSystem

We define a series of list of different coordinates over which an observation may obtain coverage. Each coordinate may be part of one or more Coordinate systems. For primary observations, each coordinate may a physical parameter. For secondary data, additional coordinate systems can be introduced.

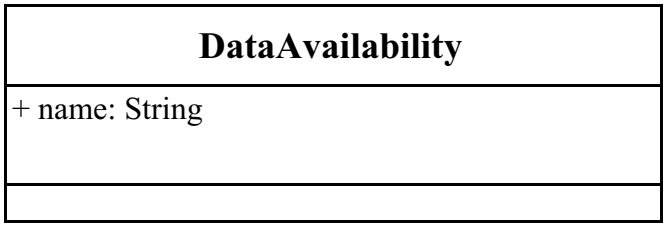
Appears in: *UOC, Solar Feature Catalogs*



DataAvailability

The data help by a provider may be stored and transferred in different manners that may effect how the system and user need to handle requests for such data. Data maybe stored on-line for realtime access, near on-line in libraries and jukeboxes that may cause a slower access times, off-line requiring human intervention to retrieve, or in other forms. The

Appears in: *Resource Registry*



DataFormat

A dataset that has been obtained may be stored or made available in multiple formats. The choice of formats may be driven by the need to reduce the volume of a dataset or the desire to make the data available in a widely readable format. A data format may or may not incorporate compression. Different compression schemes may be *lossless* or *lossy*, the latter involving the loss of information that makes the data less useful for quantitative work. The amount of loss may range in severity from minor (e.g. a small reduction in dynamic range) to major (e.g. large quality reduction for a JPEG image). Common data formats include FITS, PNG, JPEG, and MPEG.

Appears in: *Resource Registry*

DataFormat
+ formatName: String + hasCompression: String + informationLoss: integer

Dataset

A Dataset describes the a collection of data which have been grouped together based on some operational or scientific motivation. A Dataset may be composed of a single image or may be a grouping of data covering some period of time. A Dataset may also be made up of data obtained from multiple Producers with extended coverage in one or more coordinates (e.g. temporal or spectral).

Appears in: *Creator Registry, UOC*

Dataset
+ name: String + dataSetType: String

Deployment

A Deployment is the installation of a certain Assembly at a single Observatory, referring to a single location, for a defined Duration. A Deployment defines the position of a given Assembly (and hence Instrument) at a particular moment.

Appears in: *Creator Registry*

Deployment
+ name: String + deploymentType: String

Distribution

The Distribution describes, to various levels of detail, exactly how the covered volume was actually sampled in one or more coordinate axes. Except in the case of a regular sampling, the coverage over a given dimension may not be fully described by the statistical and encompassing description given by the Coverage. There may be points within the enclosed volume that were not sampled, or the sampling might have different weightings within the defined range. Examples of such Distributions include the precise definition of a field of view (including non-rectangular fields of view) or the transmission profile of a filter.

A Distribution may be presented in multiple ways, with various levels of detail or complexity. For example, a filter passband may be given as an analytical function that describes the general shape of the passband, (e.g. a Gaussian or two-cavity profile) possibly including the parameters necessary for calculating the particular instance profile. Alternatively, the passband may be given as an array of weightings at multiple points within the relative range, either quantitatively (e.g. a table of values) or descriptively (e.g. an image), possibly referred to through a link to an external source. The spatial distribution may be described as a function (e.g. circular annulus from 1-3 solar radii), as the layout of pixels over the field of view (e.g. as given by the CDELT*n*, CROTA*n*, and associated keywords in the FITS formality), or as a mask that graphically or numerically enumerates the pixel

coverage.

This class may be best implemented as a superclass with multiple subclasses each specifically describing the different representations of a distribution. The best way to divide and describe these subclasses will be found through practical implementation efforts.

Appears in: *UOC*

Distribution
+ distributionType: String + functionName: String + functionParameters: String

DistributionDefinition

A Distribution may be described by defining regions according to some common and simple region definition. Such definitions may include the description of common 2-D areas such as rectangle, ellipses, or arbitrary polygons.

Appears in: *UOC, Solar Feature Catalogs*

DistributionDefinition
+ name: String

DistributionMask

A Distribution may be described by a map of the positions which are part of the associated Distribution. This may be given in different forms, such as a mask of positions in a simple array form, or as a chain code description of the boundaries of the Distribution. These representations will presumably be stored externally, with a pointer to the location being maintained within the system.

Appears in: *UOC, Solar Feature Catalogs*

DistributionMask
+ name: String

DomainType

The data model will be applied to data coming from instruments that obtain measurements of different classes of particles. The two primary domains are the electromagnetic regime of photons and the particle regime of the protons, electrons, and ions. Different coordinate axes and measurement types may apply to the different domains, and hence need for discrimination among multiple domains.

Appears in: *Creator Registry, Resource Registry*

<<enumeration>>

DomainType

+ electromagnetic: String
+ particle: String

EnergyRegime

An instrument is restricted, by its design and physics, to operate in only a certain set of energy regimes. The boundaries among these regimes are essentially arbitrary, though the division should minimize the splitting of instruments into multiple adjacent bands. Again, the intent of this classification is less to imply a physical division but rather as a common search criteria in wide use within the domain. This classification could be used for both wavelengths for the observation of photons and for energies in the measurements of particles.

Appears in: *Creator Registry*

EnergyRegime

+ name: String
+ range: SpectralInterval
+ domain: DomainType

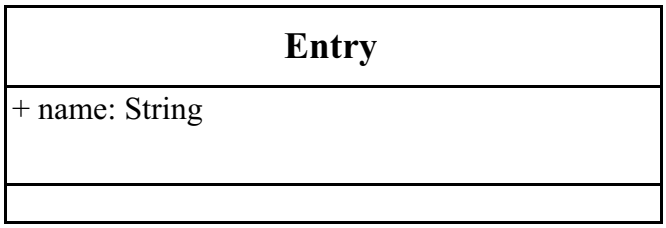
A possible division of energy regimes for the electromagnetic spectrum is the following:

Energy Regime	**Energy**		**Wavelength**		**Frequency**	
	min	*max*	*min*	*max*	*min*	*max*
Gamma Ray	50 keV			0.025 nm	1.2×10^{19} Hz	
X-Ray	0.124 keV	50 keV	0.025 nm	10 nm	3×10^{16} Hz	1.2×10^{19} Hz
Extreme Ultaviolet	0.0124 keV	0.124 keV	10 nm	100 nm	3×10^{15} Hz	3×10^{16} Hz
Ultraviolet	12.4 eV	3.8 eV	100 nm	320 nm	9×10^{14} Hz	3×10^{15} Hz
Visible	3.8 eV	1.24 eV	320nm	1000 nm	3×10^{14} Hz	9×10^{14} Hz
Near Infrared	1.24 eV	0.124 eV	1000 nm	10000 nm	3×10^{13} Hz	3×10^{14} Hz
Far Infrared	1.24×10^{-3} eV	0.124 eV	10 □	1000□	3×10^{11} Hz	3×10^{13} Hz
Radio		1.24×10^{-3} eV	1000 □			3×10^{11} Hz

Entry

A collection of descriptive elements that conveys a consistent unit of information about a solar feature as compiled in a catalog. An Entry will generally contain some combination of temporal or spatial specifications, quantitative measurements, qualitative classifications, among others. Note that a single "record" in a solar feature catalog may actually be divided into multiple entries.

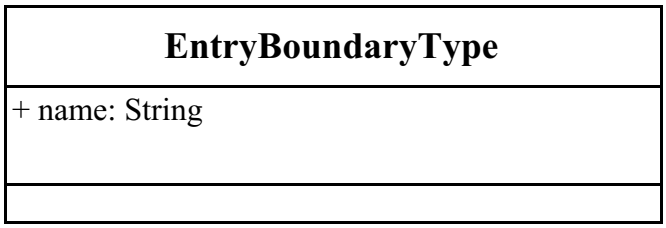
Appears in: *Solar Feature Catalog*



EntryBoundaryType

The definition of the extremes of a spectral, spatial, or temporal interval may be limited by constraints from different sources. The means by which the boundary is determined may effect how the boundaries are treated. These boundaries may be limited by observational constraints, measurement errors, or actual limits on the evolution of the feature.

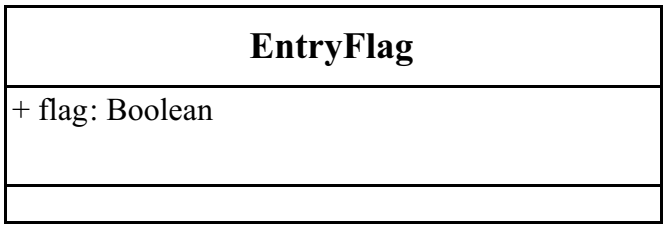
Appears in: *Solar Feature Catalog*



EntryFlag

An EntryFlag is a type of Classification for which the only two possible values are to have the flag be raised or not. A null value may also be allowed in addition to the normal Boolean values.

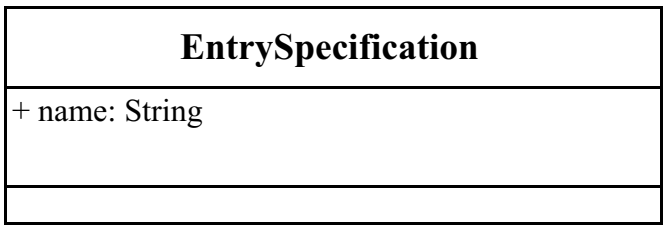
Appears in: *Solar Feature Catalog*



EntrySpecification

A generalization of the different specifications of temporal, spectral, and spatial location for an entry in a solar feature catalog.

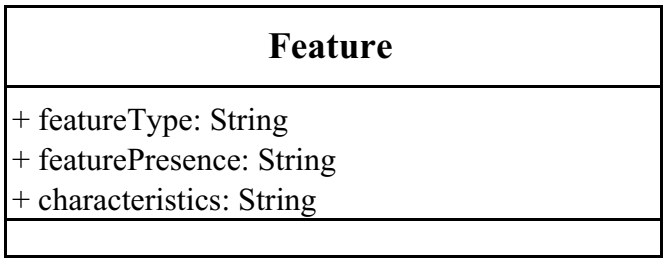
Appears in: *Solar Feature Catalog*



Feature

A Feature records the particular occurrence of some element (as defined in FeatureType) in the solar atmosphere or heliosphere. The Feature may simply be noted as having occurred at a certain time or place, or more specific characteristics may be measured using one or more DataSets.

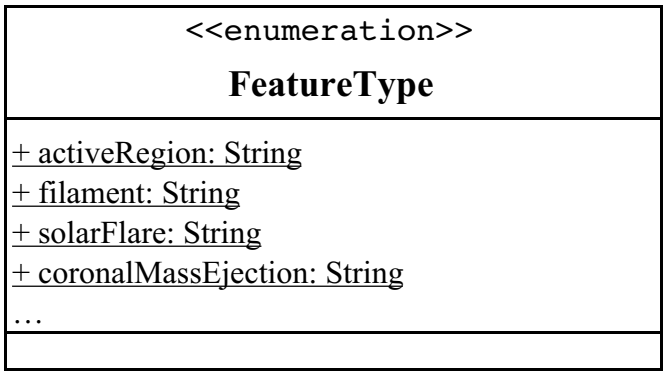
Appears in: *Solar Feature Catalogs, UOC*



FeatureType

The Feature Type is a set of the numerous solar features which have been classified in the solar atmosphere or heliosphere. The list is determined through domain specific classification.

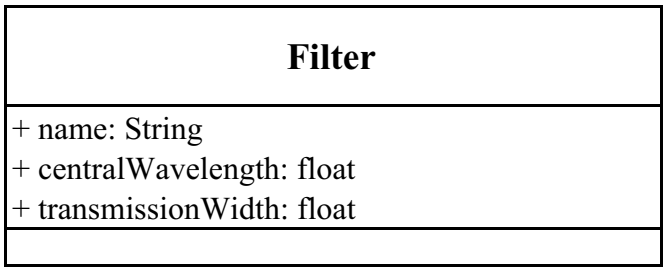
Appears in: *Solar Feature Catalogs, UOC*



Filter

It is common to obtain data using spectral filters that only allow some small portion of the spectrum to be transmitted. It is also common to refer to datasets by the common names for the Filter used to acquire the data. The name of a Filter may be related to its central wavelength or other descriptive items (e.g. filter composition – "thin aluminum"). A Filter may also be referred to by the width of the transmission peak (often called the Full Width at Half Maximum or FWHM), which determines the character of the data obtained and their suitability for different purposes (e.g. velocity measurements).

Appears in: *Creator Registry, UOC*



Generator

A Generator is an entity that utilizes one or more Packages and possibly incorporates one or more Datasets, to produce a SecondaryDataset. The Generator is generally a stable entity with a responsible organization that produces a fixed type of data for many years. A Generator may also represent a more ad hoc structure that generates a SecondaryDataset on a more occasional basis.

Appears in: *Resource Registry*

Generator
+ name: String

Instrument

An Instrument is the combination of certain essential elements to allow for the measurement of photons or particles from the Sun or heliosphere in order to provide a sampling in one or more physical observables. An Instrument acquires data in some stable manner, recording information in some limited set of modes that result in similarities among the data obtained by a single instrument. An Instrument operates in a fixed number of energy regimes, which generally do not vary for a single instrument.

Appears in: *Creator Registry*

Instrument
+ name: String + acronym: String + notoriety: integer

InstrumentType

The different types of instruments can be described according to a general classification of the different instruments utilized in solar and heliospheric physics. These classifications are defined within the domain, and may not obtain complete logical consistency. Rather, the primary use of this classification will be as selection criteria for the knowledgeable user or through the intelligence incorporated into the Consumer search tools.

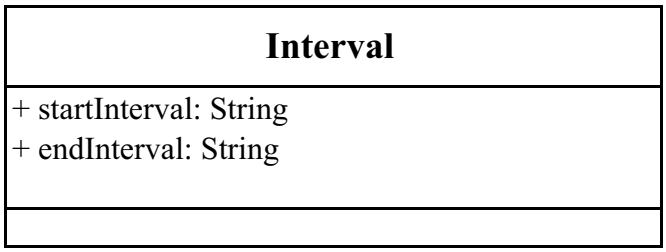
Appears in: *Creator Registry*

<<enumeration>> **InstrumentType**
+ coronagraph: String + magnetograph: String + polarimeter: String + spectrograph: String …

Interval

An Interval represents an extent in some coordinate or axis that indicates a non-finite coverage. The Interval is defined as a starting and ending point on that coordinate, though the meaning of each extreme position may be governed by the associated BoundaryType. It is also possible to define a partially "open" interval, where the limits may be variable, such as in the case of the operating interval of a still functioning instrument. The limits of the interval may also be defined with errors defining the uncertainties in the location of those limits.

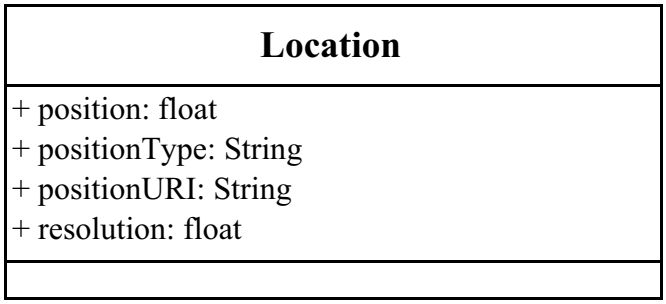
Appears in: *Resource Registry, UOC, Creator Registry*



Location

The information necessary to calculate the position, at any given moment, of an Observatory is defined as a Location. A location may be represented in multiple coordinate systems. The location may describe a single location, such as the latitude, longitude, and altitude of a fixed observatory on the Earth's surface, or a path or trajectory for an orbiting spacecraft. The Location may provide the actual coordinates or orbital elements, where appropriate, or pointers to services that can provide the position at a given time. The Location for a single observatory may be given in multiple resolutions. This may be needed in cases where a lower resolution location is suitable for most purposes (e.g. the SOHO satellite is at L1), but information on the exact position at any given time might be needed for some studies (e.g. the orbit of SOHO around L1 for helioseismological analysis).

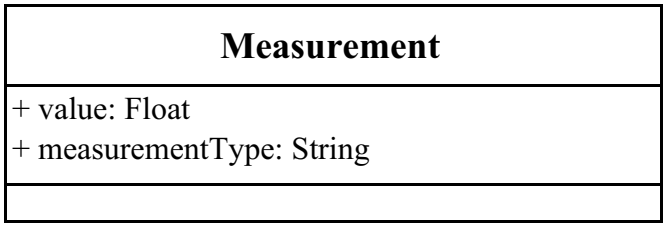
Appears in: *Creator Registry*



Measurement

A Measurement is a quantitative element of derived metadata that indicates a value obtained from the observed data. The Measurement is a measured quantity concerning a feature recorded in a solar feature catalog. Examples of measurements include velocities, morphological parameters, or peak fluxes. A Measurement may be an observational type, where the value comes directly from a single observation, or of the derived type, where the value is extracted from the analysis of multiple observations, often taking into account understanding of the specific type of feature observed.

Appears in: *Solar Feature Catalog*



Network

A series of Instruments with similar characteristics or goals can be combined into a organized entity, called a network, that operates in some coordinated manner or produces data with some common purpose. The network may manifest itself as a tight coupling among essentially identical instruments, or it may be an *ad hoc* collection of a variety of independently operated instruments. In the former case, the network may operate as a meta-instrument, producing datasets that are the result of the merging of data from multiple elements of the network, while in the latter case, the network operates more as a virtual organization, presenting data from the multiple instruments individually as distinct data products. In this model a Network is made up of multiple Deployments and exists

for some finite Duration. New Deployments may be added to the Network, but this may result in the need to create a new Network object. Some examples of Networks might be the GONG or other helioseismology networks and the Cluster spacecraft.

Appears in: *Creator Registry*

Network
+ name: String + networkType: String

Object

An Object is a classification of the different structures and features observed in the solar atmosphere and heliosphere. This list requires a common classification of the different terms used to describe the observed structures. The starting point for this list may be the OBJECT list produced as part of the SOHO project.

Appears in: *UOC, Solar Feature Catalog*

<<enumeration>> **Object**
+ coronalHole: String + activeRegion: String + coronalMassEjection: String …

ObservationMode

There are several different broad modes in which an instrument may be said to obtain data. These different classifications may distinguish among the different datasets by indicating their purpose. Again, this is primarily a domain specific list used by the experienced user to indicate data of certain known specific types, or to exclude data which were obtained in some non-standard way.

Appears in: *Creator Registry, UOC*

<<enumeration>> **ObservationMode**
+ standard: String + patrol: String + synoptic: String + calibration: String + commissioning: String …

Observatory

An Observatory is a physical entity, under the responsibility of one or more organizations, which exists in a single defined location or in some continuous series of locations. An observatory may be a ground-based observatory occupying some fixed position on the earth's surface, a mobile observation platform operating within the Earth's atmosphere (e.g. balloon or airborne observatory), or a spacecraft that carries instruments on a trajectory in near-earth space or the extended heliosphere.

In the general case, an observatory provides the resources (utilities, support, etc.) necessary for the installation of an instrument. Observatories are usually considered to be official and long-lasting entities, though in some cases an Observatory may be defined on an *ad hoc* basis to represent a temporary (e.g. eclipse observations) or informal (e.g. amateur instrument) installation.

Appears in: *Creator Registry*

Observatory
+ name: String + observatoryType: String + observatoryLocationType: String

Orbit

An Orbit is the description of the path an object takes in its motion through the solar system. Earth orbits may be described relative to the earth (whose own orbit is well known). Interplanetary orbits may be described in a coordinate system tied to the center of the solar system. Orbits may be classified according to type: Low Earth Orbit, Geosynchronous, Interplanetary, etc. An Orbit may be defined with its analytical parameters or may merely be a pointer to an external resource that is capable of calculating the position of a certain spacecraft.

Appears in: *Resource Registry*

Orbit
+ orbitType: String + orbitParameters: Float + orbitPointer: String

Ordering

The Ordering defines how a series of Measurements or Classifications may be sorted relative to each other. Measurements will generally be sorted by normal numerical rules based on values. Classifications may also be sorted numerically in some cases, though often there may be other sorting rules. Examples of such sorting include the magnetic classification of solar active regions (Alpha, Beta, Gamma, Delta), that indicates an ordering but has its own specific sorting rules. The sorting rules may be simply defined (numerical, spectral, etc.) or may need to rely on an external definition of the ordering rules.

Appears in: *Solar Feature Catalog*

Ordering
+ orderingType: String + orderingDefinition: String

Organization

An organization is an administrative entity that plays a formal role managing some resource of interest to solar physics or the EGSO system. An organization can be the owner or have (partial) responsibility for some physical structure, such as an observatory or instrument. An organization may also be responsible for a virtual grouping, such as a coordinated observation network. An

Organization may maintain ownership or responsibility of an entity for only a limited period of time. Organizations may include institutes, universities, national bodies, or private foundations.

Appears in: *Creator Registry, Resource Registry*

Organization
+ name: String + address: String + country: String + URI: String + contact: String

Package

A Package is the installation of a certain Program. Package will generally represent the different versions of the same Program, but may also represent different installations of that Program where installation specific details may be of importance.

Appears in: *Resource Registry*

Package
+ name: String + version: String + installation: String

Physical Parameter

From the flux of incoming photons or particles, an instrument may measure, or obtain data from which it is possible to derive, one or more physical parameters. The measurements of the physical conditions may be done in absolute coordinates (e.g. magnetic field strength) or may only be provide relative measures (e.g. relative velocities). The physical parameter is a list of parameters measured by solar or heliospheric instruments. A physical parameter may be directly present in the datasets, or may be available following additional reduction, possibly based on additional assumptions.

Appears in: *Creator Registry, UOC*

<<enumeration>> **PhysicalParameter**
+ flux: String + velocity: String + magneticField: String + electricField: String + acousticPower: String …

PositionEntry

A PositionEntry is a specification of a spatial position for an Entry in a solar feature catalog, where the position may also be given with additional information such as errors in its determination.

Appears in: *Solar Feature Catalog*

PositionEntry
+ name: String

Process

A Process is the execution of a Package that produces some SecondaryDataset. The Process incorporates information about the Package used, the specific parameters used in calling that package, and the effects produced by that instance of a Process.

Appears in: *Resource Registry*

Process
+ name: String + date: Date

ProcessConfiguration

The configuration utilized in executing a specific Process. The Process configuration includes those parameters that may produce significant changes (for the user) in the resulting SecondaryDataset. The same configuration may be shared among multiple Processes.

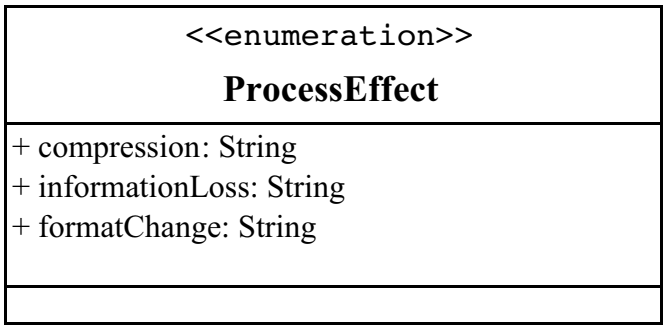
Appears in: *Resource Registry*

ProcessConfiguration
+ name: String + configurationParameters: String + configurationType: String

ProcessEffect

The primary effects produced by a Process on a dataset. Such effects may include compression, information loss, etc. These effects indicate the general changes that may be produced when applied to a primary Dataset or the difference in a SecondaryDataset with respect to the primary Dataset from which it is generated.

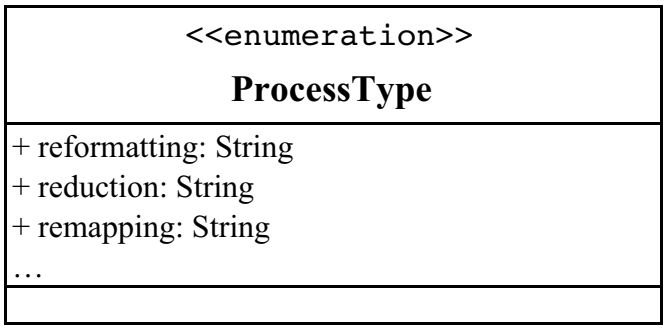
Appears in: *Resource Registry*



ProcessType

A general classification of the different types of processes that may be applied to generate secondary data. These help differentiate among different types of processes in which the user may be interested. Some examples of ProcessType include Reformatting, Reduction, Observable Extraction, and Remapping. The definition of this list will come from a domain specific understanding of the classification of processes utilized.

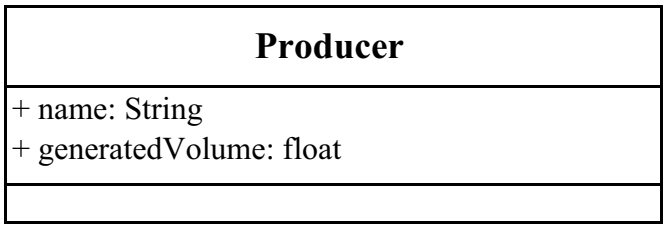
Appears in: *Resource Registry*



Producer

The Producer is the superclass that represents any entity that might be responsible for generating data. It provides a single entity with which to represent all data producers. In the data model, both Deployments and Networks can be types of Producers.

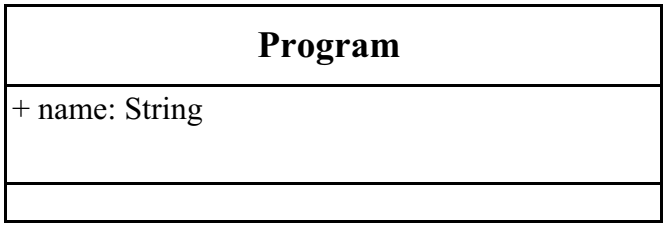
Appears in: *Creator Registry, UOC*



Program

The general software application, algorithm, or process that is used to generate or modify datasets. A program is a the general, common description of an application, without referring specifically to multiple versions or installations of the software.

Appears in: *Resource Registry, UOC*



Provider

A system that makes a resource available to consumers. A Provider is under the responsibility of one or more organizations. There may be multiple installations of the same Provider at different locations under the responsibility of separate organizations.

Appears in: *Resource Registry*

Provider
+ name: String

PublishedCatalog

A catalog that is made available by an Archive is considered to be Published Catalog. These data are provided in one or more data formats through some standardized access method encapsulated in a URI.

Appears in: *Resource Registry*

PublishedCatalog
+ name: String +

PublishedDataSet

A dataset that is made available by an Archive is considered to be Published Data, which provide the data from a given Producer covering some time period. These data are provided in one or more data formats through some standardized access method encapsulated in a URI.

Appears in: *Resource Registry*

PublishedDataSet
+ name: String + connectionQuality: float + connectionType: String + version: float

PublishedMaterial

A superset encompassing any type of data or metadata made available to the consumers in an organized fashion.

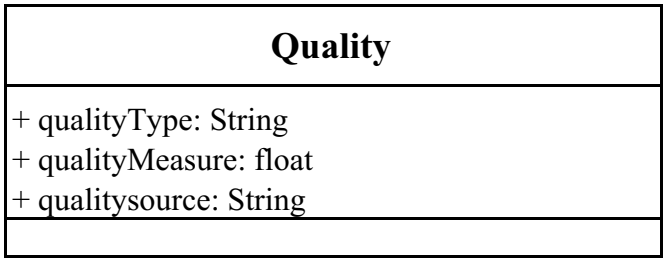
Appears in: *Resource Registry*

PublishedMaterial
+ name: String

Quality

The Quality of the data contains information that describes, either qualitatively or quantitatively the content of the data, above all with respect to its usefulness for scientific applications. Often the usage of quality terms is expressed in the negative, flagging those cases where the data is known to not adhere to certain expected conditions that facilitate scientific exploitation. The Quality may be a computed or measured value, or a human estimate. Common concepts that may fall under the Quality description include: *seeing*, describing the atmospheric conditions that may determine the spatial resolution of the data; *calibration quality*, indicating the existence or quality of the necessary calibration data (e.g. flat fields, dark current, etc.); or *orbit anomalies*, covering the possible degradations that occur when obtaining data in the South Atlantic Anomaly, or the effects of the atmosphere during the begin and end of the daylight portion of an orbit.

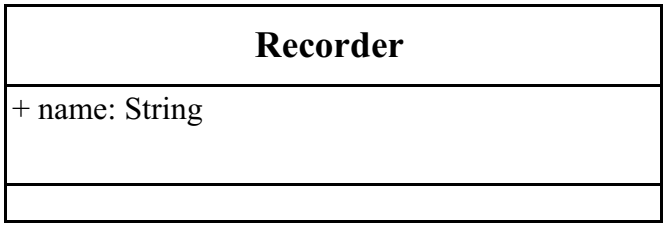
Appears in: *UOC*



Recorder

A system that gathers and stores the original raw data that comprise a primary dataset. A Recorder may be either the Deployment of an Instrument or a collection of multiple instruments forming a Network.

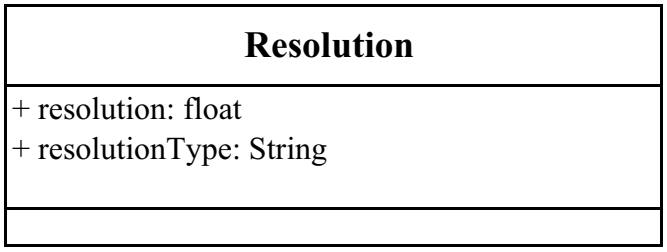
Appears in: *Resource Registry*



Resolution

The Resolution describes the level of detail that can be distinguished along a given coordinate as recorded in an associated dataset. The resolution may describe the capabilities of the instrument that acquired the data, or the limitations within a certain dataset due to external factors (e.g. atmospheric turbulence or spacecraft motion).

Appears in: *Resource Registry*



Responsibility

The operation or usage of a certain entity generally falls under the responsibility of one or more Organizations. These responsibilities can fall into several different classifications, including ownership, operational responsibility, or information contact. The responsibilities can be used to determine whom to contact regarding the use or availability of certain instruments or data. The

responsibility relationship between organization and entity can be limited to a finite duration.

Appears in: *Creator Registry*

Responsibility
+ responsibilityType: String

Sampling

In order to define a model that will be valid for the broadest range of solar observations possible, we apply statistical measures to the sampling in order to derive a series of parameters that can generically describe a solar observation.

Appears in: *UOC, Solar Feature Catalogs*

Sampling
+ start: Date + end: Date + count: integer + regularity: float + coverageFactor: float

We describe these different attributes in more detail below.

Start	The beginning of the range sampled by the observations. This may be well defined in some cases, such as the starting time of an observation, or may be slightly arbitrary, as in the case of a filter, which does not have a clean cutoff, but rather a transmission that decreases below some chosen threshold value.	
	Common FITS Keywords	Examples
	DATE-OBS **WAVEMIN**	2001-06-30T17:07:19 1195 Å

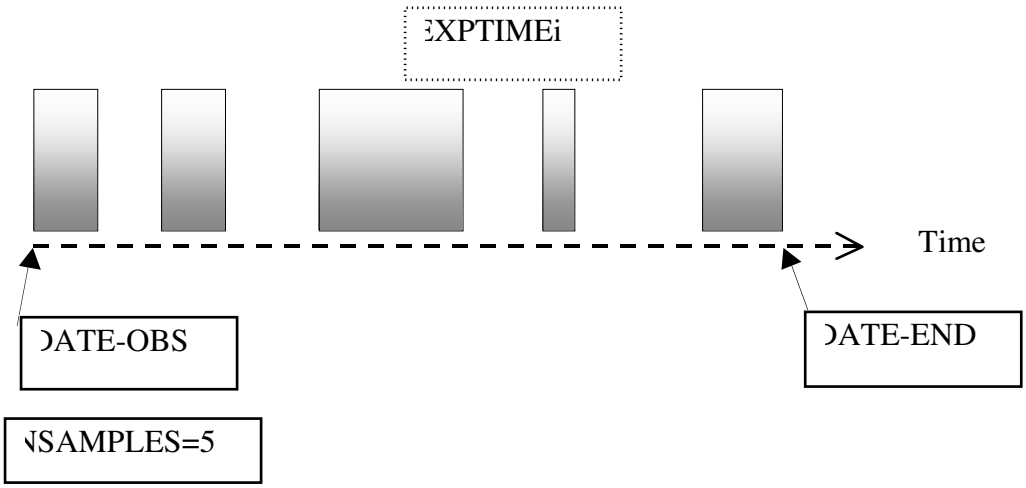
End	The end of the range sampled by the observations. As with **Start**, this may be well defined, or a more subjective decision based on instrumental knowledge.	
	Common FITS Keywords	Examples
	DATE-END **WAVEMAX**	2001-06-30T19:07:17 1243 Å

<table>
<tr><td>Count</td><td>The number of samples obtained for the corresponding range defined for an observation. This number is generally the number of physical sampling elements in the resulting data array, not the effective number of samples (e.g. the number of pixels in an image, not the total true resolution elements as constrained by the optics or the atmosphere.
For an observation that is constructed through the combination or mosaicing of multiple sub images, the Count value would refer to the number of pixels or data elements in a given axis in the combined data array that is or could be the result from such a mosaicing.
For a single sample along a certain dimension within the defined range, such as a single image or integrated intensity, the value of Count would be 1.
<table>
<tr><td>Common FITS Keywords</td><td>Common Examples</td></tr>
<tr><td>COUNT
CADENCE
NAXISn</td><td></td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td>Regularity</td><td>A measure of the uniformity of the spacing of the Count samples along a given dimension. In some cases, such as the sampling performed by an array detector (e.g. CCD), the spacing of the samples is completely even. In other cases, such as a temporal series with occasional interruptions or the extraction of only certain spectral bands from an image, the regularity may decrease.
<table>
<tr><td>Common FITS Keywords</td><td>Common Examples</td></tr>
<tr><td></td><td></td></tr>
</table>
</td></tr>
</table>

<table>
<tr><td>Coverage Factor</td><td>A measure of the overall coverage of the physical domain in the range defined by the Start and End points. The sampling of points in the domain may not provide continuous coverage over a given axis. In some cases, such as the dimensions sampled by array detectors or when the ratio of “exposure time” to “readout time” for a detector is elevated, the filling factor will be very high, denoting complete coverage. In other cases, such as an observing day plagued with numerous interruptions or down time, the filling factor will be low.
<table>
<tr><td>Common FITS Keywords</td><td>Examples</td></tr>
<tr><td></td><td></td></tr>
</table>
</td></tr>
</table>

Example Description of Temporal Sampling



NSAMPLES is the number of different spatial samples, SEXPTIMEi is the exposure time of each single "i" sample; all the "i" values of this parameters cannot be included into the Unified Catalog because NSAMPLES is not known a priori.

$$\Delta T = [DATE\_END] - [DATE\_OBS]$$

$$[COV\_FACT] = \frac{\sum SEXPTIME_i}{\Delta T}$$

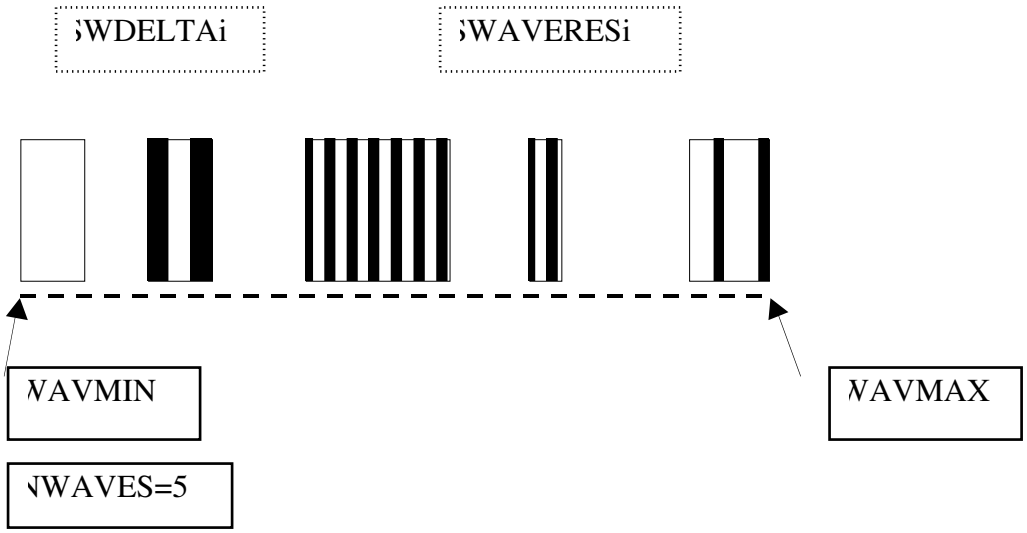
$$EXPTIME = \frac{\sum SEXPTIME_i}{NSAMPLES}$$

$$[REG\_FAC] = 100 * \sqrt{\frac{\sum (SEXPTIME_i - EXPTIME)^2}{NSAMPLES - 1}}$$

$$CADENCE = \frac{\Delta T}{NSAMPLES}$$

Note that CADENCE is a derived quantity.

Example of description of the Wavelength parameter



NWAVES is the number od different wavelength bands, SWDELTAi is the length of each single "i" wavelength bands and SWAVERESi is the spectral resolution of each single "i" wavelength bands; all the "i" values of these parameters cannot be included into the Unified Catalog because NWAVES is not known a priori.

$$[WAVMIN] - [WAVMAX]$$

$$[WCOV\_FACT] = \frac{\sum SWDELTA_i}{NWAVES}$$

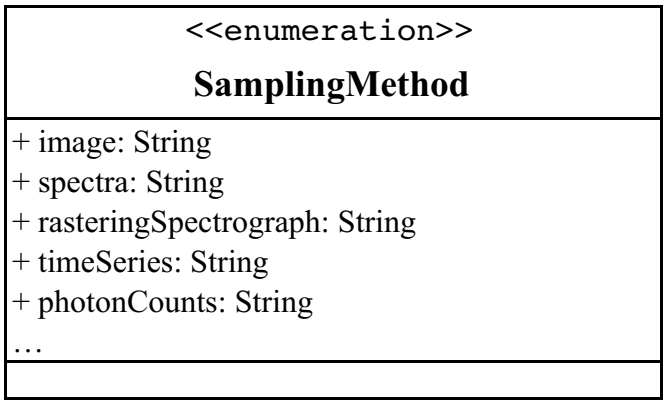
$$WAVERES = \frac{\sum SWAVERES_i}{NWAVES}$$

$$[WREG\_FAC] = 100 * \sqrt{\frac{\sum (SWAVERES_i - WAVERES)^2}{NWAVES - 1}}$$

SamplingMethod

The type of data produced by an instrument can be classified according to some broad divisions among the different general types of sampling or presentation used to generate those data. This list is based, to large part, on domain specific classifications and terminology. A single instrument may be able to generate primary data of one or more Data Types, and a dataset may encompass one more Data Types. This classification may be used by the experienced domain user to select data of only certain kinds, or the system to find data most relevant to a particular request.

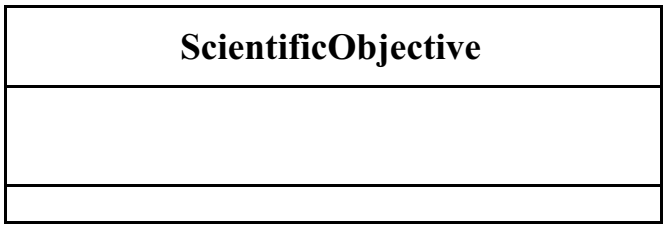
Appears in: *Creator Registry, UOC*



ScientificObjective

A description of the scientific purpose(s) that motivated the acquisition of a dataset. The ScientificObjective is a description of the scientific content of a dataset. It may describe the project, observational program, or coordinated effort that guided the definition of the observing parameters. The ScientificObjective may also describe the general type of target observed.

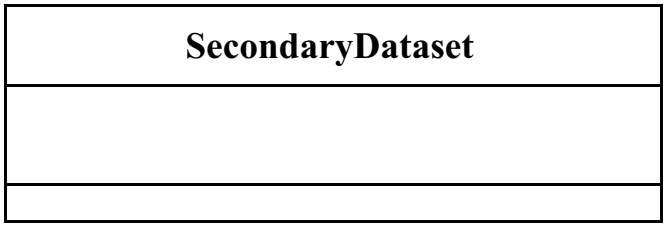
Appears in: *UOC*



SecondaryDataset

A dataset that is the result of processing (summarizing, modifying, transforming, etc.) Primary Data. A SecondaryDataset may be the combination of one or more calibrated or processed datasets.

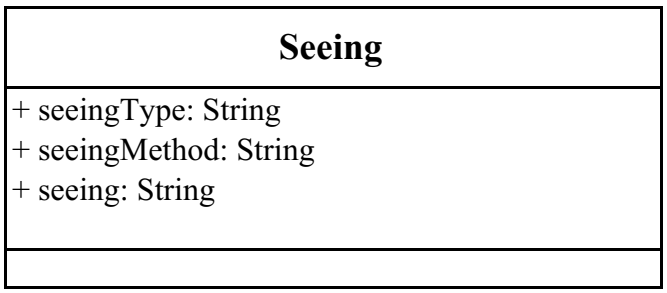
Appears in: *Resource Registry, UOC*



Seeing

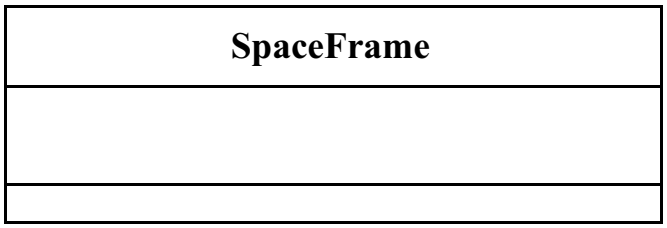
A description of the atmospheric or environmental conditions (e.g. for space-based instruments) that effect the quality or usability of a specific dataset. The Seeing can be measured qualitatively or quantitatively.

Appears in: *Resource Registry*



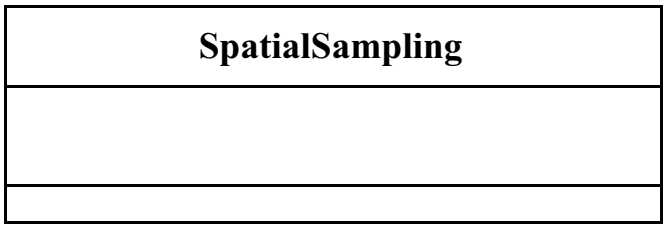
SpaceFrame

The spatial reference frame in which a coordinate system is defined.
Appears in: *Creator Registry, UOC*



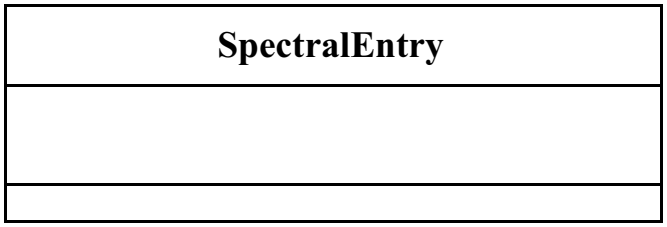
SpatialSampling

The specific description of a Sampling in one or more spatial dimensions.
Appears in: *Resource Registry*



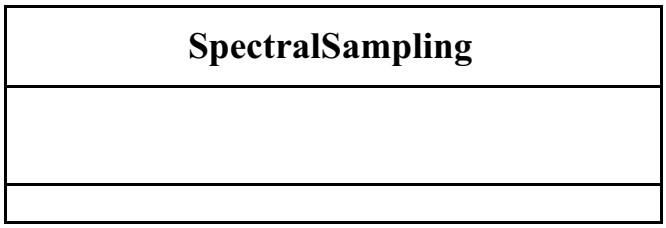
SpectralEntry

The information concerning a solar feature's spectral properties as stored in a record in a Solar Feature Catalog. May describe the specific spectral range in which the feature was observed or specific spectral signatures for a given feature.
Appears in: *Solar Feature Catalog*



SpectralSampling

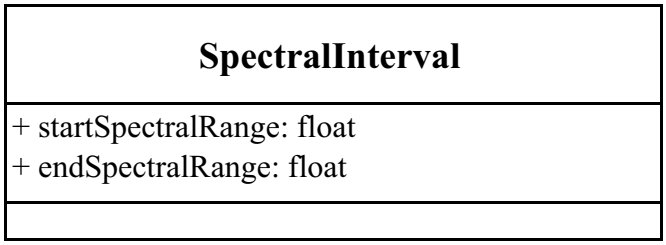
The specific description of a Sampling in the spectral dimension.
Appears in: *Resource Registry*



SpectralInterval

This is a range in wavelength/frequency/energy, covering both a starting and ending point in wavelength. If the starting and ending points are not included, the range in considered to be boundless is that direction.

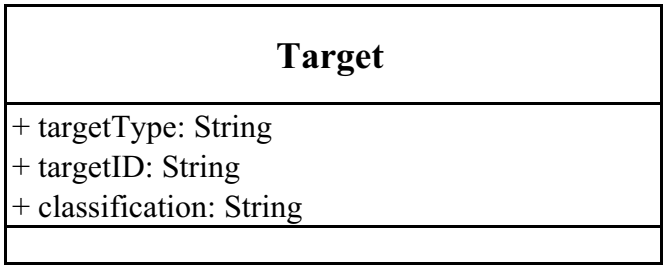
Appears in: *UOC, Creator Registry, Resource Registry*



Target

A dataset is sometimes obtained through the pointing to or observation of a specific feature in the solar atmosphere or heliosphere. This may be a named or classified feature, such as a numbered active region, or an anonymous structure, such as a prominence or CME. The Target describes the physical entity, both spatial (e.g. sunspot) or temporal (e.g. five-minute oscillations or flares) that were the focus of a particular observation. The target may be associated with features recorded in solar feature catalogs, or a general list of known solar structures.

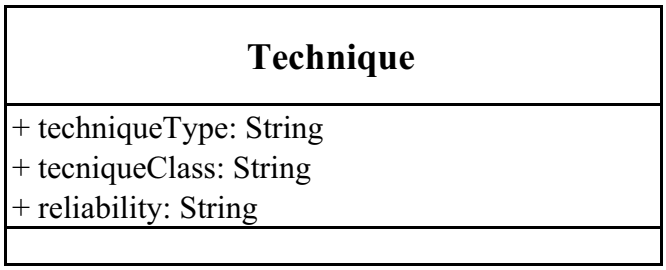
Appears in: *UOC, Solar Feature Catalog*



Technique

The identification of solar features or events, and hence the generation of a feature catalog, may be performed using a variety of techniques, both automated and relying on human discrimination. Different techniques may have certain advantages or disadvantages in the recognition of certain objects. Hence, the technique utilized may be used to discriminate among multiple catalogs of the same type of feature depending on the user's search criteria.

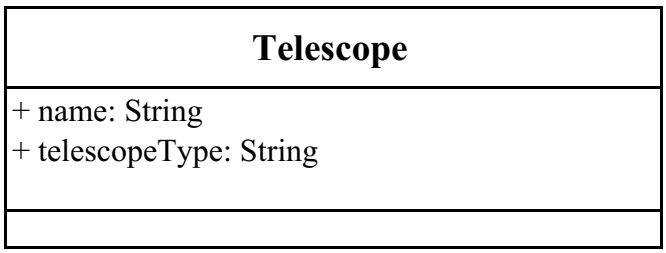
Appears in: *Solar Feature Catalog*



Telescope

A Telescope is a system for collecting photons or particles to be fed to an instrument for additional discrimination and acquisition. A Telescope does not have any provision for recording data. Not all Instruments need to have telescopes, either explicitly or implicitly.

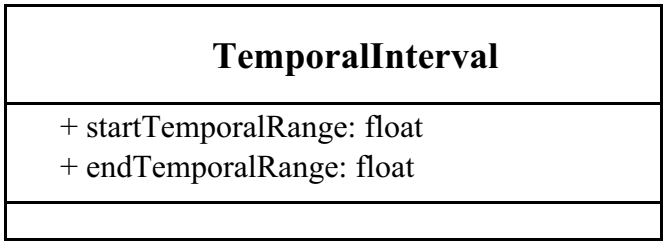
Appears in: *Creator Registry*



TemporalInterval

The description of a temporal range, with a starting and ending time. If the starting and ending points are not included, the range in considered to be boundless is that direction. A TemporalInterval may represent the durations for which an entity or grouping may exist. A TemporalInterval can be an overall lifetime, defining the absolute beginning and end of an entity or relationship, or a extended interruption, in cases which the availability or functioning of an entity is not continuous.

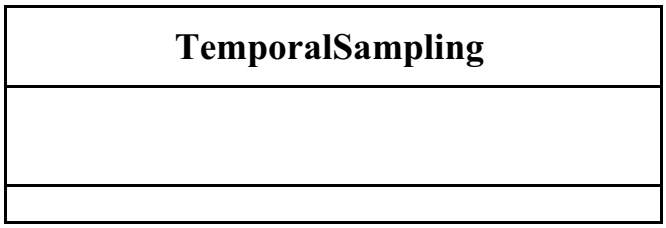
Appears in: *UOC, Creator Registry, Resource Registry*



TemporalSampling

The specific description of a Sampling in the temporal dimension.

Appears in: *Resource Registry, UOC*



TimeEntry

A TimeEntry is the description of a specific temporal element for a solar feature recorded in a Feature Catalog. The TimeEntry may describe a single moment or some finite temporal interval

during the feature's lifetime. All features must have at least one TimeEntry.

Appears in: *Solar Feature Catalog*

TimeEntry

TimeEntryType

A TimeEntry given for a solar feature may have different meanings. A time may be given as an observed time, describing a solar feature at a specific moment in it lifetime as defined by the observations themselves. A time may also be given as a derived time, identifying a specific moment in a feature's lifetime such as the starting, ending, or peak time, as derived from multiple observations.

Appears in: *Solar Feature Catalog*

TimeEntryType

TimeFrame

The TimeFrame or Epoch for which a coordinate system is defined.

Appears in: *Resource Registry, UOC*

TimeFrame

URI

The Uniform Resource Identifier indicates an address and method by which a resource may be accessed. The URI may refer to the locator for a single resource, or the address by which multiple resources may be accessed.

Appears in: *Resource Registry*

URI
+ address: String + protocol: String

Appendix C. Sample Catalog Representations

8.1 EIT

Broker Summarized View

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	2002/09/25		1321	0.12	0.87
	00:00	23:59:59			
Spatial	*x:* -1342” *y:* -1287”	*x:* 1540” *y:* 1366”	*x:* 1024 *y:* 1024	0.77	0.73
Spectral	170 Å	340 Å	1	0.2	1

User Data Browsing View

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	2002/09/25		131	0.37	0.77
	07:19:35	09:47:18			
Spatial	*x:* -1342” *y:* -1287”	*x:* 1540” *y:* 1366”	*x:* 1024 *y:* 1024	0.89	0.73
Spectral	170 Å	340 Å	1	0.2	1

Single Image View

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	2002/09/25		1	1	1
	07:34:45	07:34:46			
Spatial	*x:* -1042” *y:* -1217”	*x:* 1006” *y:* 931”	*x:* 1024 *y:* 1024	0.97	1.0
Spectral	170 Å	340 Å	1	0.2	1

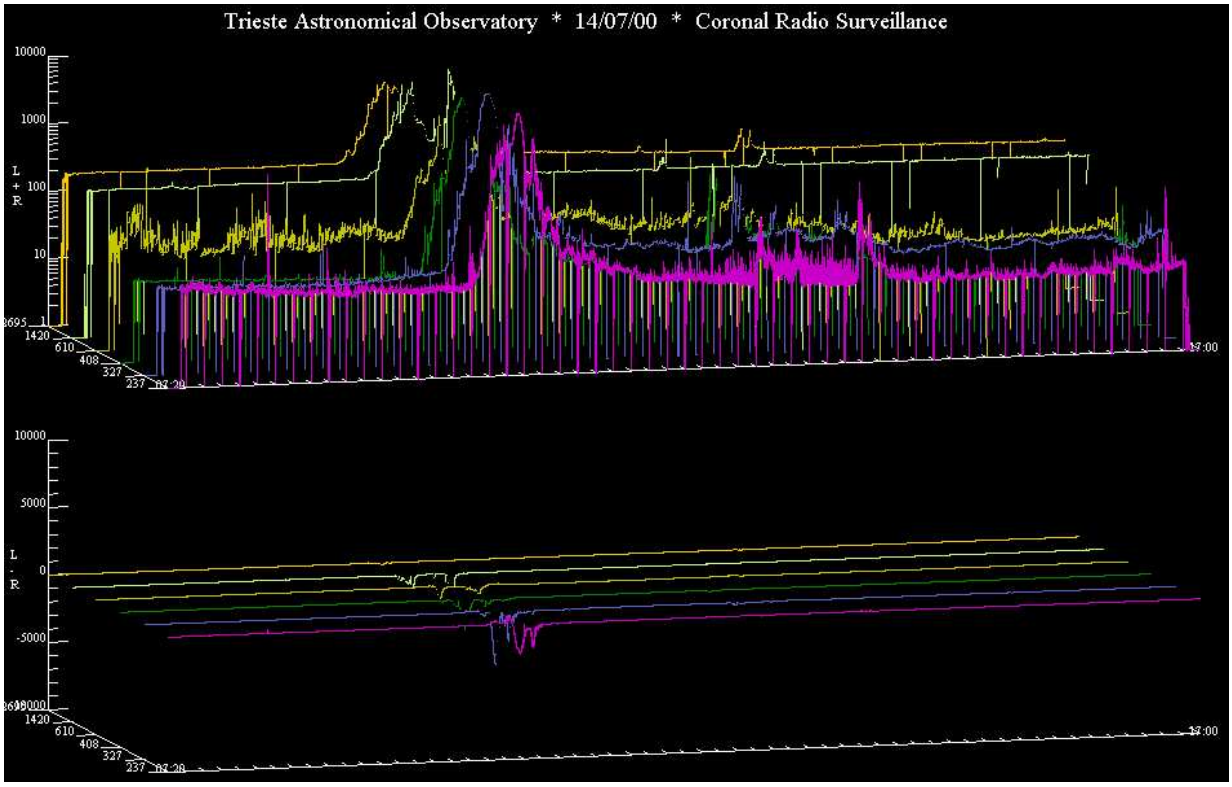
8.2 Solar Radio Observations

Observatory**: Trieste Solar Radio System, Basovizza**
Instrument: Decimetric and metric parabolic antennas
Type of observation: Multichannel 237-2695 MHz, circular polarization, no spatial resolution

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	2003/01/27 11:20:00	2003/01/27 11:30:00	600000	1.0	1.0
Spatial	-	-	-	-	-
Spectral	237 MHz 1,264947 m	2695 MHz 0,11124 m	6	?	?

Polarization: circular left, circular right

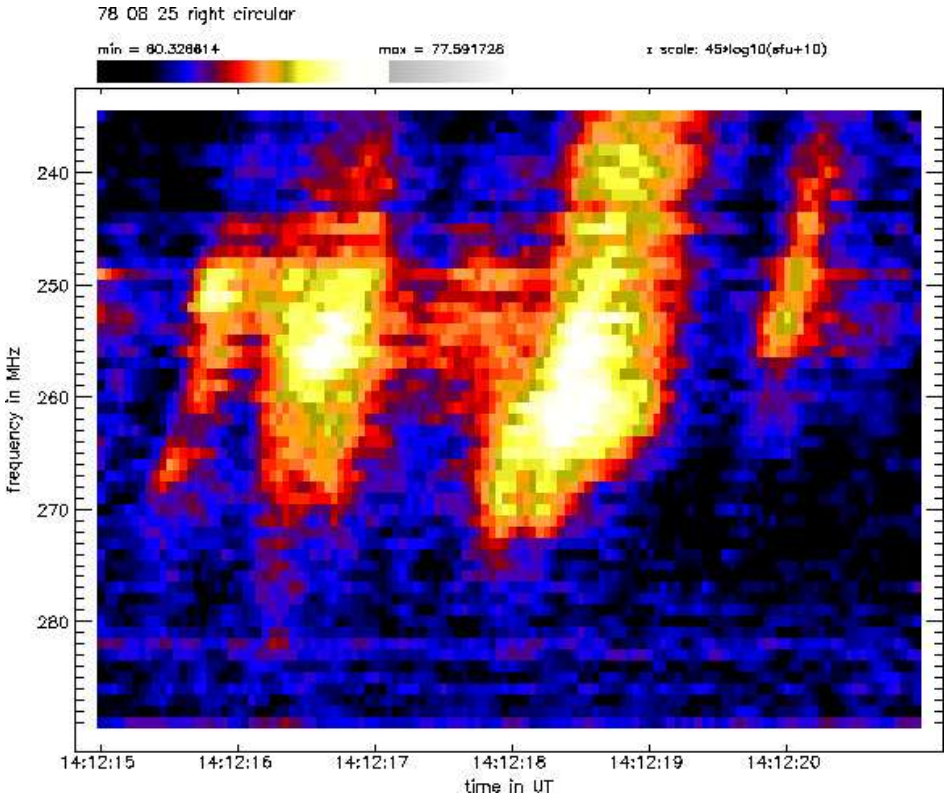
Sample image:



Observatory: **Institute of Astronomy ETH Zurich**
Instrument: "Phoenix" Parabolic radio antennas
Type of observation: Spectrograms 0.1-4 GHz, circular polarization, no spatial resolution

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	2002/5/17 12:30:00	2002/5/17 12:45:00	9000	1.0	1.0
Spatial	-	-	-	-	-
Spectral	112 MHz	3970 MHz	200	1.0?	<1.0

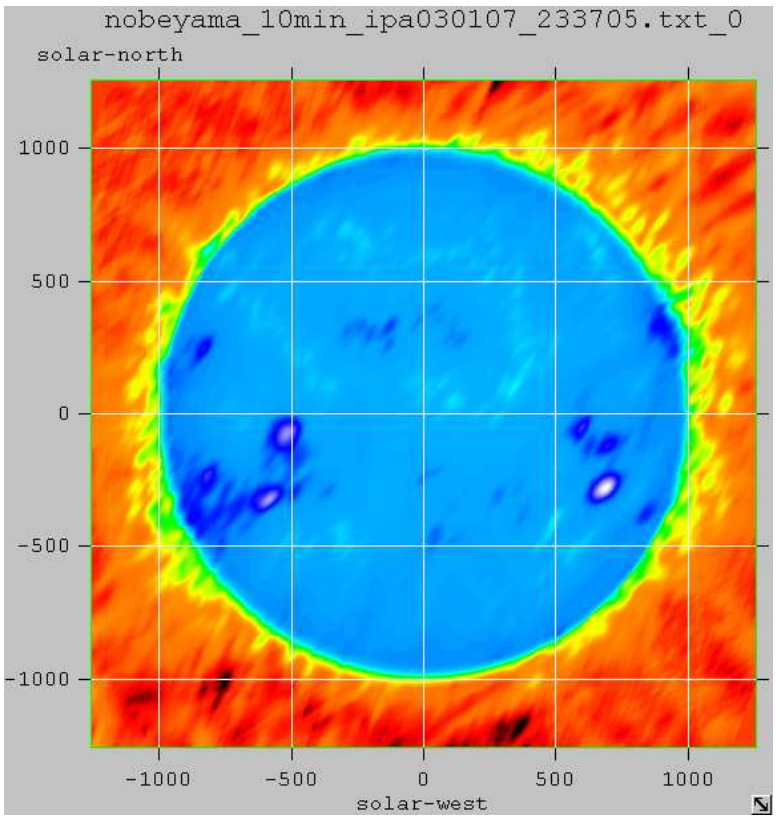
Sample image:



Observatory: **Nobeyama**
Instrument: Radioheliograph - array of 84 parabolic antennas
Type of observation: 17GHz, circular polarization L+R, spatial resolution

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	2003-01-04 00:00:02.555	2003-01-04 00:00:07.555	1	-	-
Spatial	-1259.69 arcsec	+1254.78 arcsec	512	1.0	1.0
Spectral	17 GHz	17 GHz	1	-	-

Sample image:



8.3 UVCS

UVCS Single Pointing View

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	01-JAN-1997 04:49:33	01-JAN-1997 04:58:49	5	0.90	
Spatial	*r:* 270 *R :* 1.40 R_o	*r:* 270 *R:*1.40 R_o	*x:* 1 *y:* 120	1	
Spectral	1195.3Å	1243.4 Å	256	0.34	0.4

UVCS Single FITS File View

	Start	End	Number of Samples	Coverage Factor	Regularity
Temporal	01-JAN-1997 04:49:33	01-JAN-1997 06:37:15	57	0.88	
Spatial	*r:* 270 *R :* 1.40 R_o	*r:* 270 *R:*3.09 R_o	*x:* 7 *y:* 120	0.2	
Spectral	1195.3Å	1243.4 Å	256	0.34	0.4