The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

310/1749-25

ICTP-COST-USNSWP-CAWSES-INAF-INFN
International Advanced School
on
Space Weather
2-19 May 2006

Solar Data Handling

Robert BENTLEY
UCL Department of Space and Climate Physics
Mullard Space Science Laboratory
Hombury St. Mary
Dorking
Surrey RH5 6NT
U.K.

Solar data Handling

Rob Bentley
University College London (UCL)
Mullard Space Science Laboratory

Trieste, 3 May 2006
International Advanced School on Space Weather

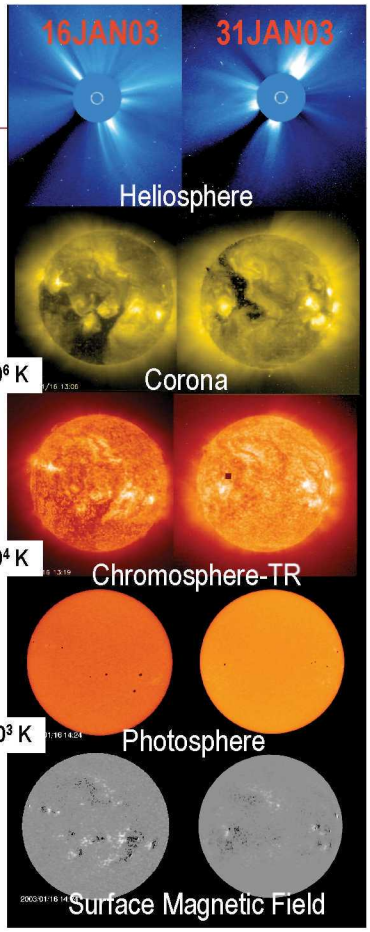
European Grid of Solar Observations

Outline

- Nature of Observations
- Data management issues
  - Types of data
  - Storage of data
  - Access to data (current)
- Standards
  - Temporal and spatial coordinates
  - File Formats
- Types of Metadata
- Data Models

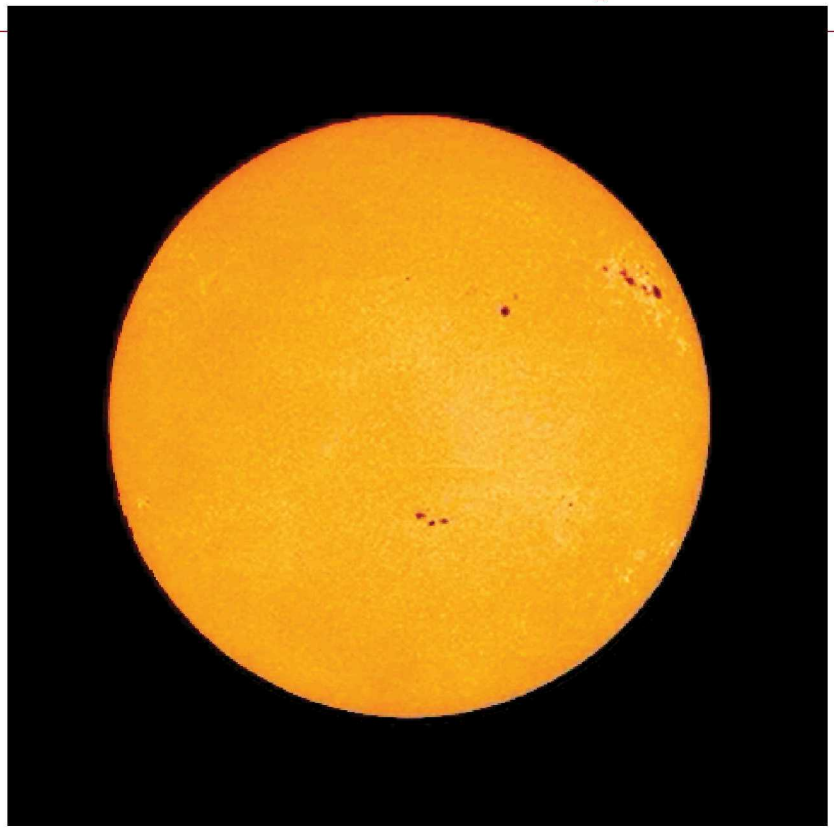
Nature of solar observations

- The appearance of the Sun changes dramatically with wavelength
  - Emissions originate from different layers in the atmosphere and different physical phenomena
- For a complete picture we need to use as wide a range of observations as possible
  - Mixture of types of observations
  - Different wavelengths and time intervals
  - Mixture of observations from space- and ground-based platforms
- Increasing desire to study problems that span communities
  - Desire relevant to Space Weather, climate physics, planetary physics, astrophysics
  - Identifying observations across community boundaries and then retrieving them are major problems



European Grid of Solar Observations

Sun at different wavelengths



Types of Data

- Types of data/observation:
  - single values
  - spectra (includes and dispersed quantity)
  - images
  - compound - e.g. spectroscopic (scan) in images (CDS, FCS)
  - multiple related parameter (heliospheric)
- Observation usually repeated over time (time series)
- Instrument/observatory location:
  - Remote sensed
    - Most solar observations made remotely, mostly on Sun-Earth line
    - SOHO, TRACE, RHESSI
  - In situ
    - Plasma physics observations of actual plasma, location important
    - ACE, GEOTAIL, Ulysses, ISEE, etc.; Cassini, Mars & Venus Express, etc
  - New observatories breaking the rule
    - Located away from Su-Earth line; location important…
    - STEREO, Solar Orbiter, LWS Sentinels - break the mold...

egso

Storage of data

European Grid of Solar Observations

- Many types of data stored in different types of container
  - Legacy data includes different formats – e.g. photographic plates, etc.
  - Most common form now is electronic data stored as files
- Many different file formats
  - FITS, CDF, HDF, ASCII/text, proprietary, etc. (more later)
  - Different formats each suited to different types of data
  - Format used depends on the type of data, institution and community
- Some differences in the files:
  - Time interval stored in each file
    - Single image or spectra in a file
    - Time series of simple parameters (hour, day, week, year...)
    - Multiple observations in a file
      - Yohkoh – orbit, similar format for 4 different instruments
      - TRACE – hour, thousands of JPEG compressed images
  - Levels of processing of the data
    - Reformatted into files, differences on how much processing done
    - Solar - "raw"+calibration; astrophysics, etc. - processed
    - Whether data calibrated has implications of when analyzing the data

Storage of data

- Longevity of the data is a major issue
  - What format should be used and what level of processing?
    - Cannot be sure that in 10 years time can still access the data
    - Proprietary formats can cause problems…
  - Expertise in the data has relatively short lifetime
    - Proper documentation of calibrations etc. algorithms essential
  - Supplied read routines good, but will they always work…

egso

Access to Data

European Grid of Solar Observations

- Electronic data not always easy to get at!
- Introduction of the internet and availability of cheap storage made huge difference, but not uniformly applied
- Variety of ways of accessing the data
  - Data can be on-line, near-line, off-line
  - Several different protocols: FTP, HTTP, cgi-bin...
  - Data in a range of file formats with no standardization of file names or directory structure
  - Some data only available through controlled access
    - Need to know exact address of file to access it
    - Access controlled by the PI
- Data providers have different resources available to them
  - Small/large providers, archives, etc.
  - Ability to service data requests varies greatly
- Use of the virtual observatory paradigm to try to address many of the problems
  - Solar data grids (SDG) and solar virtual observatories (SVO)

Standards

- Discuss standards in several areas
  - Time systems
  - Systems of Spatial Coordinates
    - Solar
    - Space Plasma
  - File Formats
  - Translation between standards is often possible, but does not always make sense
- Why are standards important?
  - It is important to be able to understand each other…
  - Because Space Weather crosses community boundaries
  - Because there have not been proper standards in the past and need greater uniformity in the future
    - Data that are VO compliant are much easier to handle!!!

egso

Time Standards

European Grid of Solar Observations

- International Atomic Time (TAI)
  - statistical timescale based on a large number of atomic clocks. (second defined in terms of cycles or radiation in atomic transition of Cesium 133)
- Universal Time (UT)
  - counted from 0 hours at midnight, with unit of duration the mean solar day, defined to be as uniform as possible despite variations in the rotation of the Earth.
- Coordinated Universal Time (UTC)
  - differs from TAI by an integral number of seconds. UTC is kept within 0.9 seconds of UT1 (form of UT) by the introduction of one-second steps to UTC, the "leap second."
- Other standards
  - Dynamical Time (TDT and TDB), Geocentric Coordinate Time (TCG), Barycentric Coordinate Time (TCB) and Sidereal Time.
  - See: http://tycho.usno.navy.mil/systime.html

Ways of Expressing Time

- Regular Date/time
  -
- Julian Date
  - Julian Day Number
    - Count of days elapsed since Greenwich mean noon on 1 January 4713 B.C., Julian proleptic calendar
    - The Julian Date is the Julian day number followed by the fraction of the day elapsed since the preceding noon.
  - Modified Julian Date (MJD)
    - Defined as MJD = JD - 2400000.5. An MJD day thus begins at midnight, civil date.
  - Julian dates can be expressed in UT , TAI, TDT, etc. – for precise applications the timestandard should be specified
    - For example: MJD 49135.3824 TAI.
- Times of several solar datasets expressed in days since certain epoch, e.g. start of 1-Jan-1979 & msod, or similar
- Many other formats used to express time
  - For instruments working at high time resolution (msec accuracy), some formats not accurate enough
  - Format very similar – differences irritating…

egso

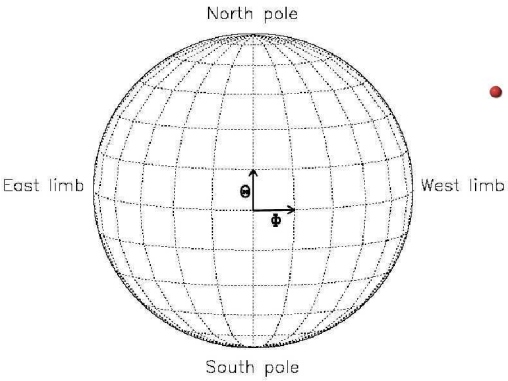
Solar Coordinates Systems

European Grid of Solar Observations

- There has not really been a standard set of spatial coordinates used in solar physics
- W. Thompson (GSFC) tried to clarify the issue when SOHO was being planned - published as paper in 2005.
- The principle coordinate sets are defined by Thompson (2005) as
  - Heliographic coordinates
    - Position of features on the Sun expressed in latitude and longitude
  - Heliocentric coordinates
    - True spatial position of feature expressed in physical units
    - System normally used when handling images as pixels arrays
  - Helioprojective coordinates
    - Equivalent to heliocentric coordinates but with the physical distance parameters replaced by angles
- Reference:

Thompson, W.T., "Coordinate systems for solar image data", *Astronomy & Astrophysics*, ??, 2005.

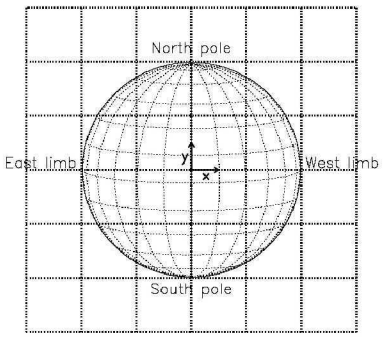
(Solar) Heliographic Coordinates

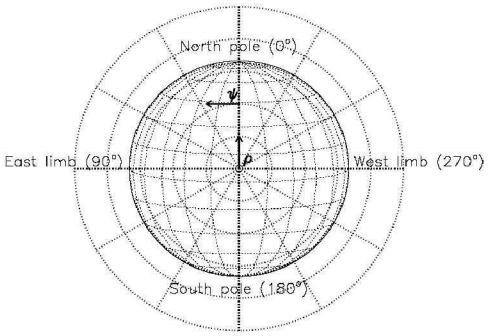


- Expresses the latitude Θ and longitude Φ of a feature on the solar surface
  - Can be extended to 3 dimensions by adding the radial distance r from the center of the Sun
- There are two basic variations of the heliographic systems
  - Both use the same solar rotational axis, but differ in the definition of longitude.
  - Stonyhurst heliographic
    - Origin of the coordinate system is at the intersection of the solar equator and the central meridian as seen from Earth.
  - Carrington heliographic
    - The coordinate system rotates at an approximation to the mean solar rotational rate (27.2753 days), as originally used by Carrington (1863).

European Grid of Solar Observations

(Solar) Heliocentric Coordinates





- Express the true spatial position of a feature in physical units from the center of the Sun
- There are two basic variations of the heliocentric systems
  - Each consists of three mutually perpendicular axes which together form a right-handed coordinate system defined relative to that observer.
  - Heliocentric-Cartesian Coordinates
    - This is a true (x; y; z) Cartesian coordinate system; the z axis is defined to be parallel to the observer-Sun line, pointing toward the observer
  - Heliocentric-radial coordinates
    - Share the same z axis as heliocentric-cartesian coordinates, but replace (x; y) with (ρ; ψ)

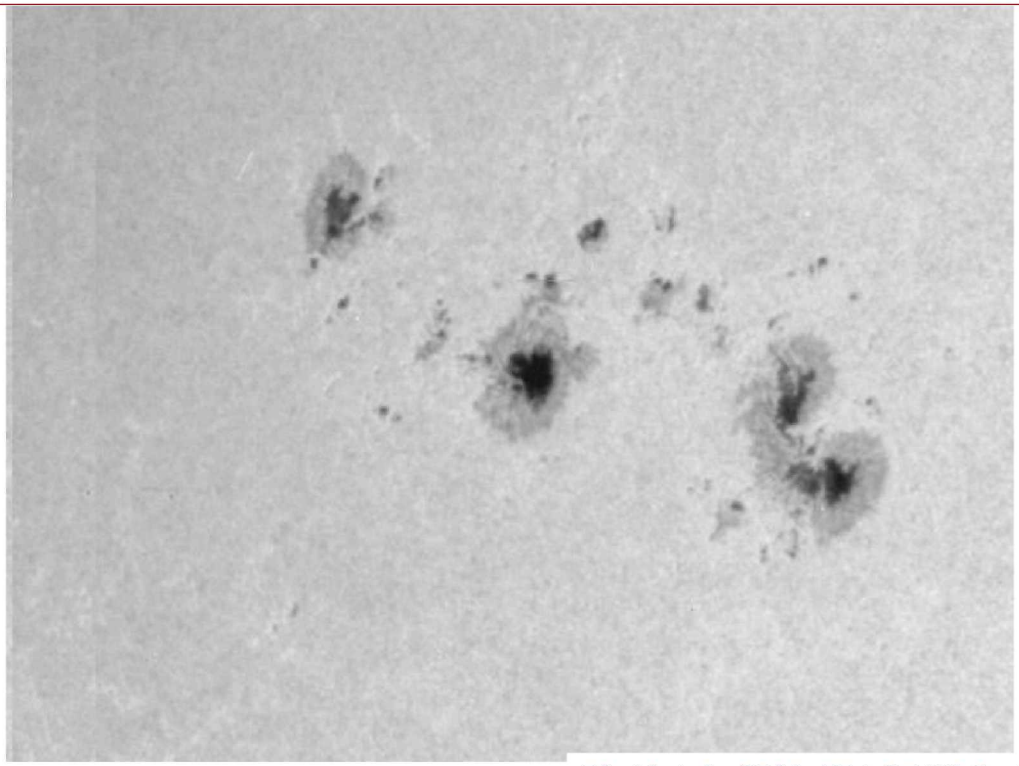
(Solar) Helioprojective Coordinates

- Data taken from a single perspective can only approximate true heliocentric coordinates
- A more precise rendition of coordinates should recognize that observations are projected against the celestial sphere
  - Known as helioprojective coordinates
  - Mimics heliocentric coordinates but replaces physical distances with angles
    - One-to-one correspondence of parameters
  - When observer is on the Earth, can be called geocentric coordinates
- Two types of helioprojective coordinates
  - Helioprojective-Cartesian Coordinates
    - Distance parameters x and y are replaced with angles θ_x and θ_y, where θ_x is the longitude and θ_y is the latitude
  - Helioprojective-Radial Coordinates
    - The impact parameter ρ is replaced by θ_ρ
  - *Approximation with heliocentric holds close to the Sun, where angles small*

egso

Tracking single Heliographic longitude

European Grid of Solar Observations



12-18 July 2001, TRACE White Light

Space Physics Coordinate Systems

- Many of the quantities measured in space physics are represented numerically by a set of components whose values depend on the coordinate system used
  - vectors (e.g. position, velocity, electric and magnetic fields, electric currents) and tensors (e.g. pressure).
- No single coordinate system which can serve all purposes
- Two principle coordinate sets are defined by Hapgood (1992) as:
  - Geocentric systems
    - These have one of their principal axes defined with respect to some natural feature of the Earth or its motion around the Sun.
  - Heliocentric systems
    - These have one of their principal axes defined with respect to some natural feature of the Sun.
- Reference:

Hapgood, M. A., "Space physics coordinate transformations: A user guide", *Planetary and Space Science*, **40**, 711-717, 1992.

egso

(Space Physics) Geocentric systems

European Grid of Solar Observations

- All Geocentric systems have their origin at the centre of the Earth and one of their principal axes defined with respect to some natural feature of the Earth or its motion around the Sun
- Three Geocentric coordinate sets:
  - Systems based on the Earth's rotation axis
  - Systems based on the Earth-Sun line
  - Systems based on the dipole axis of the Earth's magnetic field
- Systems based on the Earth's rotation axis
  - Geographic (GEO)
    - Z axis parallel to the Earth's rotation axis (positive to the North); X axis towards the intersection of the Equator and the Greenwich Meridian.
    - *Convenient for specifying the location of ground stations and ground-based experiments.*
  - Geocentric equatorial inertial (GEI)
    - Z axis parallel to the Earth's rotation axis (positive to the North); X axis towards the First Point of Aries.
    - *Convenient for specifying the orbits (and hence location) of Earth-orbiting spacecraft.*

(Space Physics) Geocentric systems - cont.

- **Systems based on the Earth-Sun line**
  - **Geocentric solar ecliptic (GSE)**
    - **X axis along Earth-Sun line; Z axis perpendicular to the plane of the Earth's orbit around the Sun (positive North).**
    - *Convenient for specifying magnetospheric boundaries and widely adopted as the system for representing vector quantities in space physics databases.*
  - **Geocentric solar magnetospheric (GSM)**
    - **X axis along Earth-Sun line; Z axis is the projection of the Earth's magnetic dipole axis (positive North) on to the plane perpendicular to the X axis.**
    - *Best system to use when studying the effects of interplanetary magnetic field components (e.g. Bz) on magnetospheric and ionospheric phenomena.*
- **Systems based on the dipole axis of the Earth's magnetic field**
  - **Solar magnetic (SM)**
    - **Z axis parallel to the Earth's magnetic dipole axis (positive North); Y axis perpendicular to the plane containing the dipole axis and the Earth-Sun line.**
    - *Preferred system for defining magnetic local time in the outer magnetosphere.*
  - **Geomagnetic (MAG)**
    - **Z axis parallel to the Earth's magnetic dipole axis (positive North); Y axis is the intersection between the Earth's equator and the geographic meridian 90° East of the meridan containing the dipole axis.**

egso

(Space Physics) Heliocentric systems

European Grid of Solar Observations

- **All Heliocentric systems have their origin at the centre of the Sun and one of their principal axes defined with respect to some natural feature of the Sun.**
- **Systems based on the Sun's rotation axis**
  - **Heliocentric Earth equatorial (HEEQ)**
    - **Z axis parallel to the Sun's rotation axis (positive to the North); X axis towards the intersection of the solar equator and the solar central meridian as seen from the Earth.**
    - **HEEQ coordinates are closely related to the Stonyhurst heliographic system**
- **Systems based on the ecliptic**
  - **Heliocentric Earth ecliptic (HEE)**
    - **Z axis perpendicular to the plane of the Earth's orbit around the Sun (positive North); X axis towards the Earth**
  - **Heliocentric Aries ecliptic (HAE)**
    - **Z axis perpendicular to the plane of the Earth's orbit around the Sun (positive North); X axis towards Aries**
  - *The HEE and HAE systems are both sometimes known as heliocentric solar ecliptic (HSE)*

General File Standards

- Flexible Image Transport System (FITS)
  - Formatted storage of 1D spectra, 2D images and 3D+ image cubes and tabular data.
  - Used in astrophysics and solar physics
  - Endorsed by NASA and the IAU; maintained by FITS Support Office at NASA/GSFC – see http://fits.gsfc.nasa.gov/
- Common Data Format (CDF)
  - CDF is a file format that facilitates the storage and retrieval of multi-dimensional scientific data
  - Used for space plasma data, etc.
  - Product of the National Space Science Data Center (NSSDC) – see http://nssdc.gsfc.nasa.gov/cdf/
- Hierarchical Data Format (HDF)
  - HDF is a multi-object file format that facilitates the transfer of various types of data between machines and operating systems
  - Used in astrophysics
  - Product of the National Center for Supercomputing Applications (NCSA) – see http://hdf.ncsa.uiuc.edu
- Other standards include NetCDF, HDF-EOS, etc.

egso

Examining a Format

European Grid of Solar Observations

- Even though file formats differ, contents often very "similar"
- FITS increasingly common in solar data
  - File has "Header" and "Data" sections
  - "Header" describes the data
    - Contents depend on nature of the data
    - Time of observation, image dimensions, exposure, wavelength, pointing, etc.
- FITS standard is loose – too "flexible"
  - Provides storage standard, but little standardization
  - "Data" part can contain different things (under extensions)
  - Requirements on what should be included in "Header" not strict

Metadata

- The information in the FITS Header is metadata
  - Metadata – data that describes data
- Descriptive metadata is the key to data and often the problem
  - Should tell user everything about the data
  - But, the contents and quality vary
    - Information often incomplete, sometimes missing completely
- Different standards for parameters
  - Different time formats
  - Different spatial coordinate systems
    - For image pointing and location of observing platform
  - Different ways of expressing qualifying quantities
    - Sampling – pixel size of imager, wavelength bin of spectrometer, etc.
    - Resolution – spatial, temporal, spectral
      - different to sampling, defined by basic instrument performance

European Grid of Solar Observations

Metadata and Data Models

There are many types of metadata:

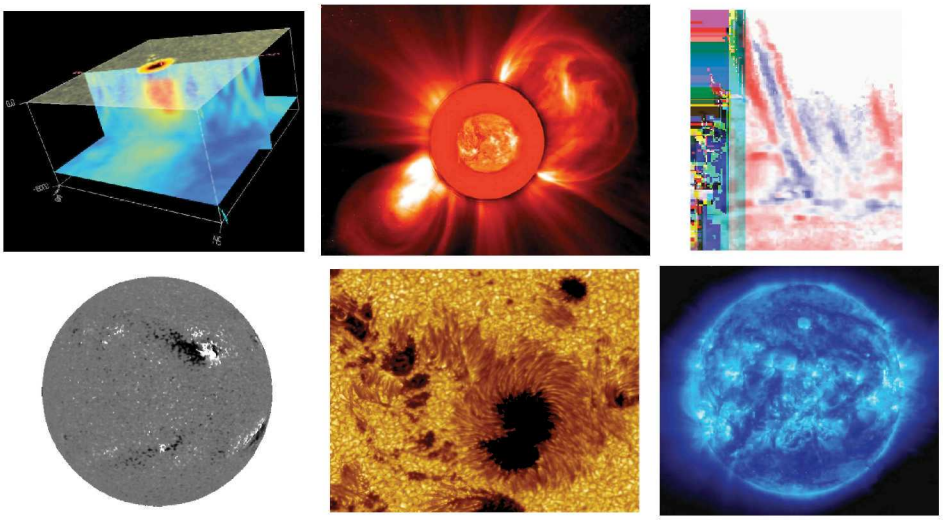
- Observational Metadata
  - Related to the way in which a given piece of Data was obtained or processed
- Derived Metadata
  - Extracted from Primary Data through subsequent analysis or processing
- Structural Metadata
  - The organizational characteristics of the data including how they related to other data
- Administrative Metadata
  - The characteristics related to access and management of Data or other Resources
- Data Model
  - The relationship amongst the metadata elements

Observational Metadata

European Grid of Solar Observations

Observational Metadata – *related to the way in which a given piece of* ***Data*** *was obtained or processed*

The Zoo of Solar Observations



Observational Metadata

European Grid of Solar Observations

- Solar data are described by many parameters related to their acquisition and subsequent processing
- No common standard for representing solar observations
  - Quality, quantity and representation of metadata varies widely
- All measurements are made on photons, particles or waves, describing their energy, direction or location, time of arrival, and polarization, etc.
- Each instrument performs these measurements differently:
  - Sampling (imager, spectrograph, photon counting)
  - Resolution (spatial, temporal, spectral)
- Operational Parameters also important:
  - Observer, Principal Investigator
  - Observational Target or Scientific Goal

The description of a series of observations is called an **Observing Catalogue**

Derived Metadata

Derived Metadata – *extracted from **Primary Data** through subsequent analysis or processing*

- Much information useful for queries currently remains locked away in the data themselves.
  - Where were the sunspots?
  - When were there flares?
  - How strong was the spectral line?
- This information must be extracted from the obtained observations.
- The derived metadata are additional catalogs that are used in the search and analysis process (SEC, SFC in EGSO).

Derived Metadata

European Grid of Solar Observations

- Some derived metadata catalogues already exists and are in wide use:
  - NOAA Solar Active Region List
  - GOES X-Ray Flare Catalog
- Generally simple, very little automated extraction.
- In EGSO, WP5 generated new, richer catalogues through the use of feature recognition techniques
  - Systematic determination of location of sunspots, filaments and active regions from synoptic solar images
- Wealth of extracted information requires a new organization and precise definitions
  - Best way to trade information between organizations
  - Area where small amount of effort to standardize can pay off

Structural Metadata

Structural Metadata – *the organizational characteristics of the* ***Data*** *including how it relates to other* ***Data***

- Structural metadata allows description of actions that can be performed on data and metadata elements.
  - the verbs to the metadata's nouns
- Allows for use of software repositories and processing services to perform actions on data in a well-defined manner.

European Grid of Solar Observations

Administrative Metadata

Administrative Metadata – *the characteristics related to methods of access and management of* ***Data or other Resources***

Purpose:

- To Locate and Access Resources within the System
  - ⇒ Resource Registry
- To Maintain User Identity, Control Resource Access
  - ⇒ Authentication and Authorization

Concerns:

- Interoperability with other Grid or Virtual Observatory projects
- Integration of Non-Data Resources
- Management of Administrative Metadata across multiple Broker instances

Data Models

- EGSO Developed a detailed Data Model when designing its data system
- Space Physics has also produced a data model
  - SPASE Dictionary
- Data modelling not an exact science
  - Based on opinion, no absolute definitions
  - Helps focus ideas on how to do things – necessary evil…
  - Needs to be implemented flexibly!!
- An organized structure of metadata is sometime called an ontology
  - Not practical to use a single ontology to describe all solar data
  - Build up smaller ontologies of subsets and create and over-spanning ontology to link them.

European Grid of Solar Observations

EGSO Metadata Class Description

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.

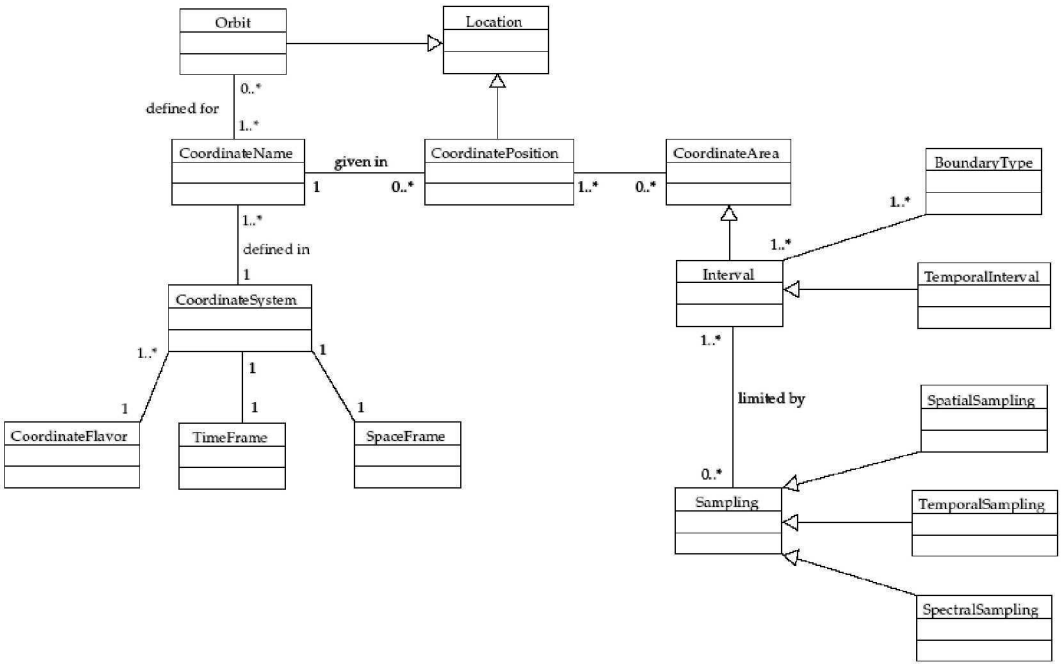


Figure 2: General Metadata Class Model

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.

EGSO Instrument Class Data Model

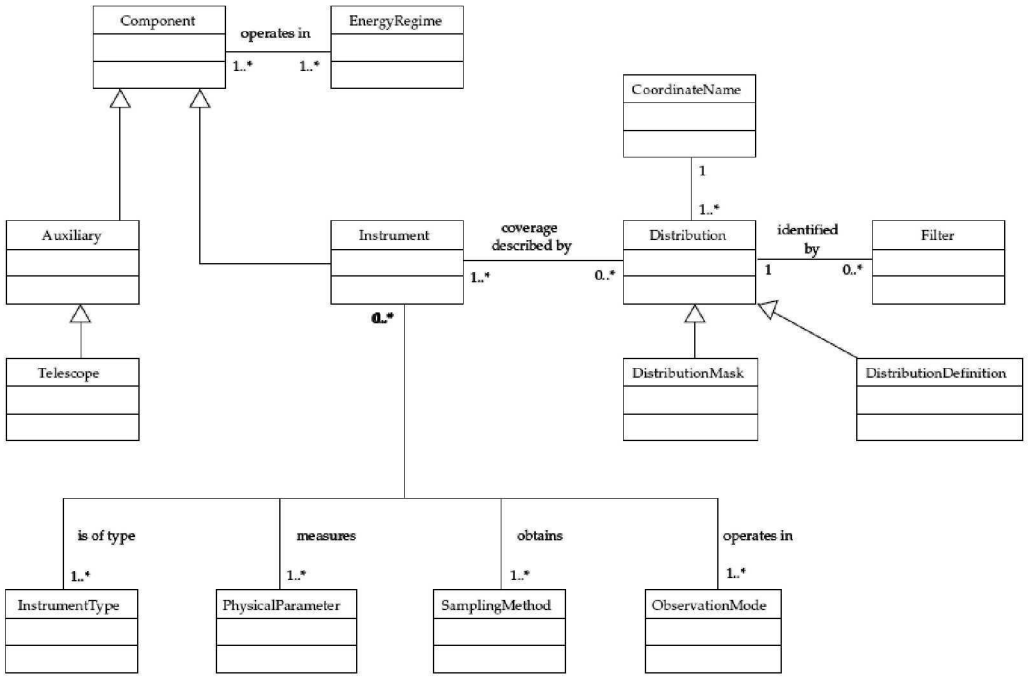


Figure 3: Instrument Class Data Model

European Grid of Solar Observations

Search Criteria: Observations

- **Search based on the basic parameters that describe the data:**
  - Time/date
  - Observation type: photons, particles, waves; images, spectra, etc; wavelength, energy range
  - Observation target/domain: solar surface, heliosphere, etc...
  - Observing location, pointing/orientation
  - Observing platform
- **List subjective and not exhaustive**
  - Differing opinions on how to describe the data
  - Source of many problems…

Search Criteria: Derived Metadata

- Search criteria for derived metadata:
  - date/time
  - events parameters (size, intensity, location)
  - feature parameters (size, location)
  - value of indices - e.g. Kp, DST, etc.
- Other possibilities depend on derived data available!!!

European Grid of Solar Observations