The Heliophysics Data Environment, Virtual Observatories, NSSDC, and SPASE
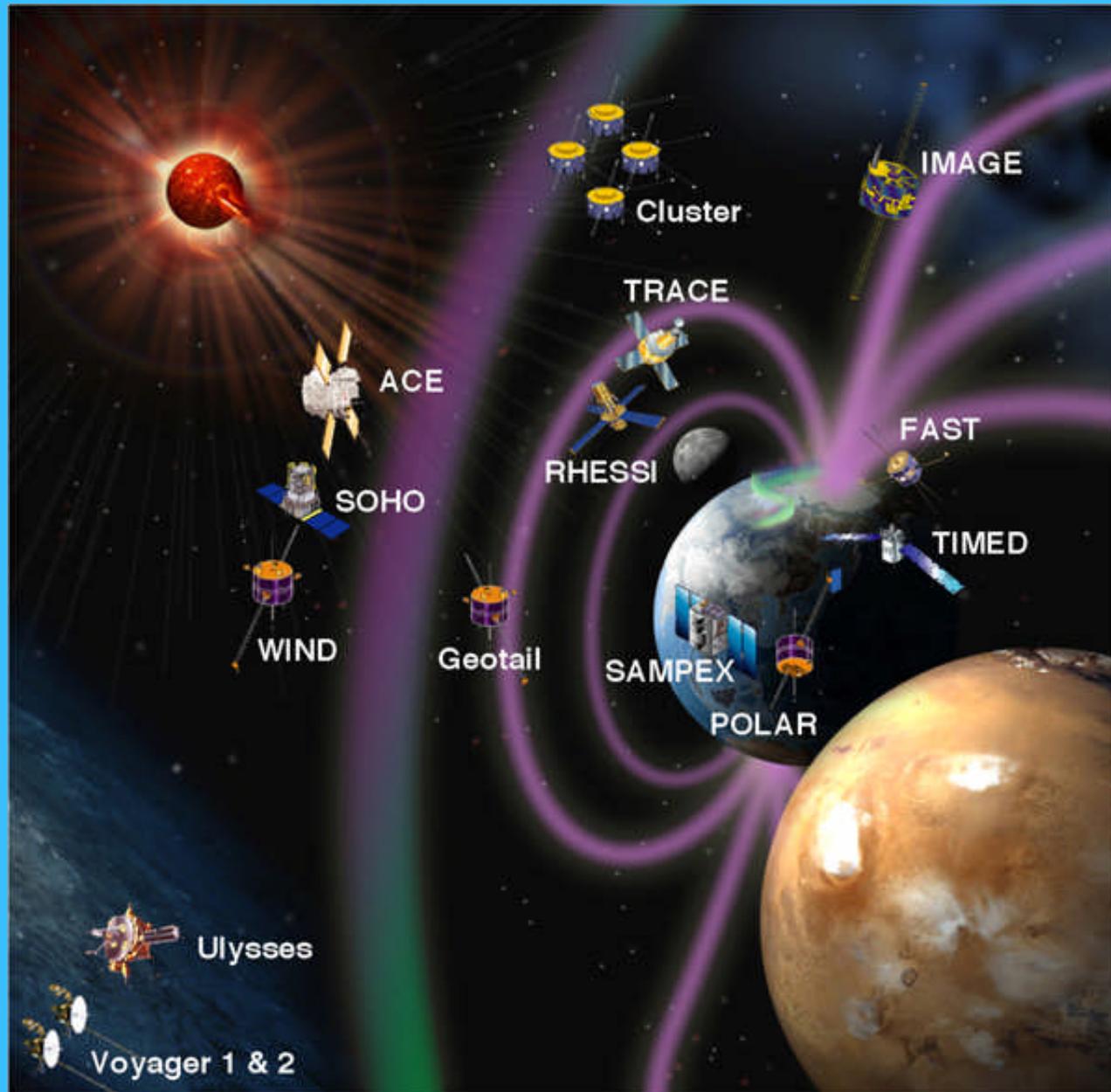
James Thieman, Edwin Grayzeck, Aaron Roberts, Todd King

Presentation at the 22nd International CODATA Conference
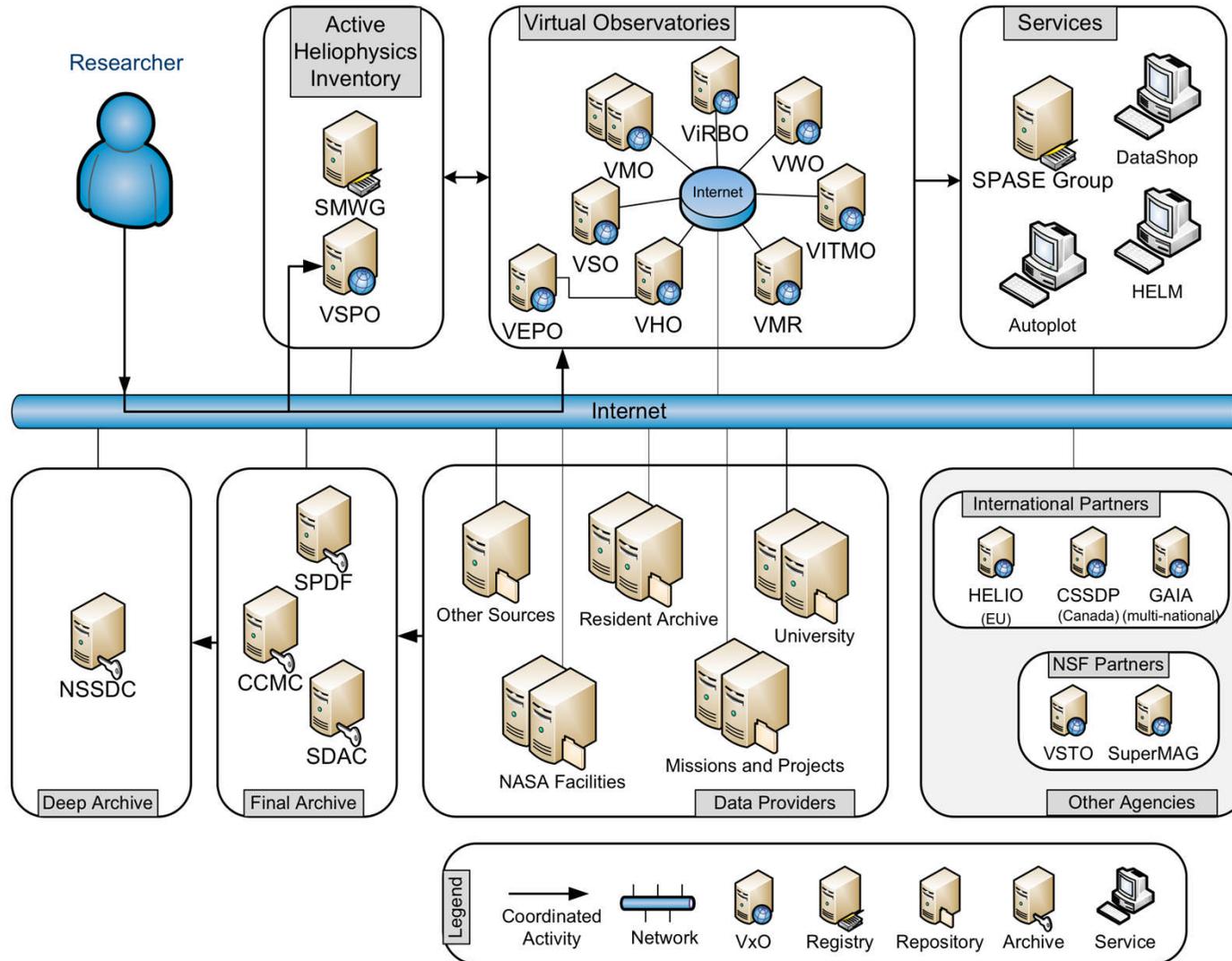October 26, 2010

OUTLINE

- **Heliophysics Data Environment**

- **Heliophysics Virtual Observatories (VO's)**

- **SPASE – VO Shared Metadata Approach**

- **SPASE Data Model**

- **Data Search within VO's and SPASE**

- **NSSDC and SPASE**

- **NSSDC as a World Data Center**

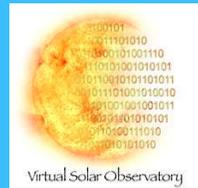- **Trusted Digital Repositories**

Heliophysics Great Observatory

Heliophysics Data Environment

Heliophysics Virtual Observatories (VOs)

NASA-Funded

- VSO - Virtual Solar Observatory
- VSPO - Virtual Space Physics Observatory
- VMO - Virtual Magnetospheric Observatory
- VITMO - Virtual Ionosphere, Thermosphere, Mesosphere Observatory
- VHO - Virtual Heliophysics Observatory
- ViRBO - Virtual Radiation Belt Observatory
- VEPO - Virtual Energetic Particle Observatory
- VWO - Virtual Wave Observatory
- VMR - Virtual Model Repository

Non-NASA-Funded

- CAA - Cluster Active Archive
- CDPP - Centre de Données de la Physique des Plasmas
- CSSDP - Canadian Space Science Data Portal
- EGSO - European Grid of Solar Observations
- GAIA - Global Auroral Imaging Access
- VSTO - Virtual Solar Terrestrial Observatory
- ??
- ??

Centre de Données de la Physique des Plasmas
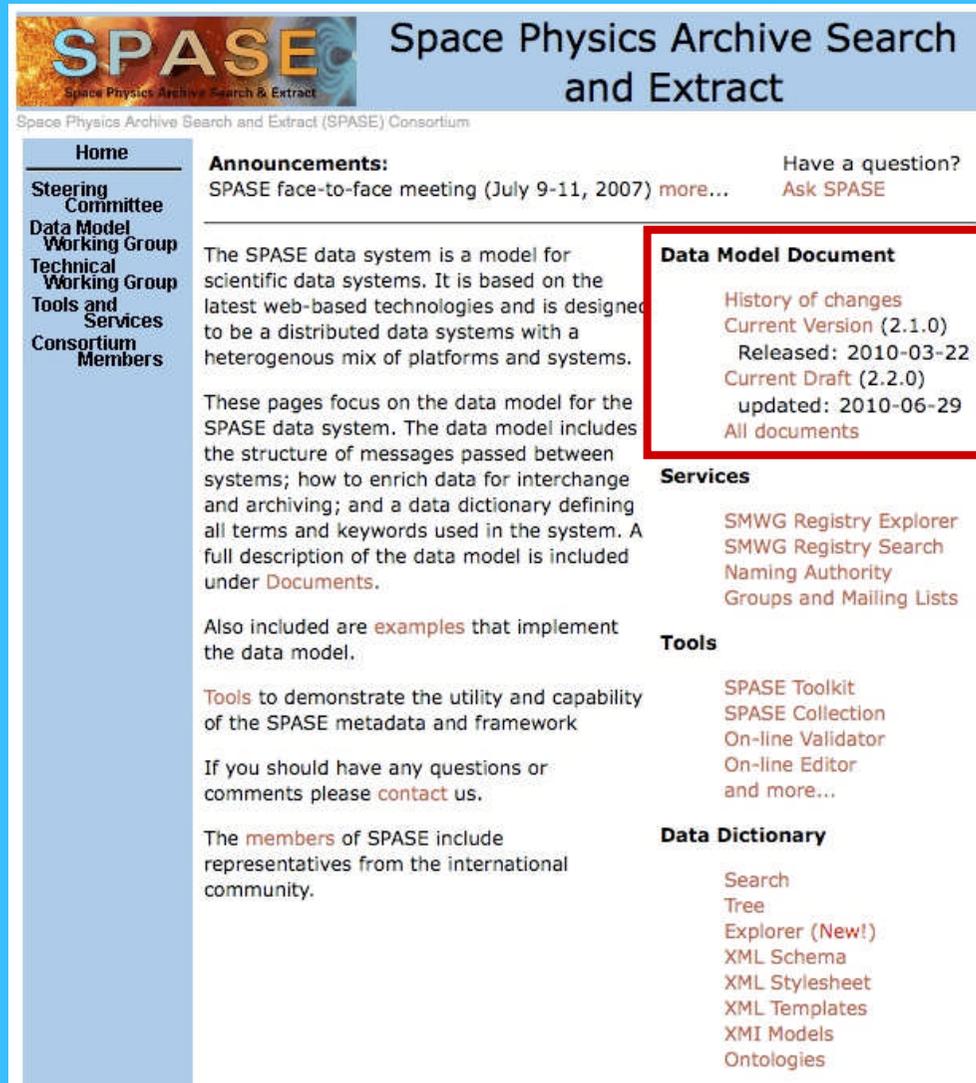Plasma Physics Data Centre

What is SPASE ?

The **Space Physics Archive Search and Extract (SPASE)** effort is a heliophysics community based project with the goals of:

- Facilitating data search and retrieval across the Space and Solar Physics data environment with a common metadata language
- Defining and maintaining a standard Data Model for Space and Solar Physics interoperability, especially among Heliophysics Virtual Observatories
- Demonstrating the Model's viability
- Providing tools and services to assist SPASE users
- Working with other groups for other heliophysics data management and services coordination as needed

The Space Physics Archive Search and Extract (SPASE) effort is implemented by the SPASE Working Group composed of representatives of the heliophysics data community

The SPASE Working Group is the only international group supporting global data management for Solar and Space Physics

ACCESS TO THE SPASE DATA MODEL

Version 2.2 of the Data Model is imminent and will be "frozen" for community usage.

http://spase-group.org

DATA MODEL DOCUMENT

A Space and Solar Physics Data Model

from the SPASE Consortium

Version: 2.1.0

Release Date: 2009-03-19

Document Generated: 2010-Mar-22

6. Examples

As an example let us describe a person using SPASE metadata. This person is "John Smith" from Smith Foundation. While the SPASE data model is implementation neutral, XML representation is preferred. This example uses the SPASE XML form.

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<Spase>
   <Person>
      <ResourceID>spase://person/jsmith@smith.org</ResourceID>
      <PersonName>John Smith</PersonName>
      <OrganizationName>Smith Foundation</OrganizationName>
      <Address>1 Main St., Smithville, MA</Address>
      <Email>jsmith@smith.org</Email>
      <PhoneNumber>1-800-555-1212</PhoneNumber>
   </Person>
</Spase>
```

4. The Data Model presented hierarchically

The taxonomy tree shows the inter-relationship of e
picture" view of the SPASE data model. This taxon
element are contained in the data dictionary.

Notes: Occurence specifications are enclosed in par
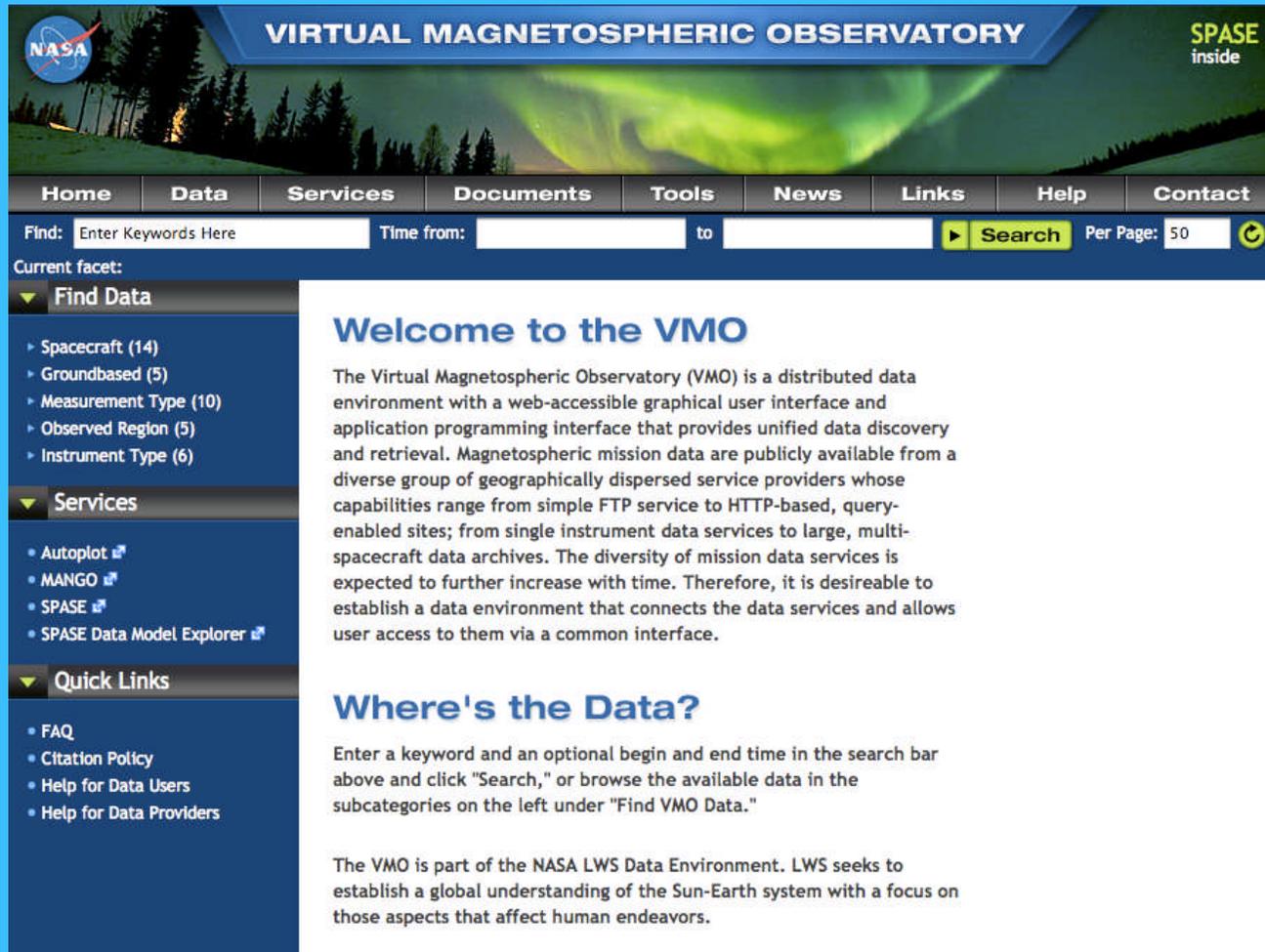more, + = 1 or more

```
+ Spase (1)
|    + Version (1)
|    + Catalog (* of A)
|    |    + Resource ID (1)
|    |    + Resource Header (1)
|    |    |    + Resource Name (1)
|    |    |    + Alternate Name (*)
|    |    |    + Release Date (1)
|    |    |    + Expiration Date (0)
|    |    |    + Description (1)
|    |    |    + Acknowledgement (0)
|    |    |    + Contact (+)
|    |    |    |    + Person ID (1)
|    |    |    |    + Role (+)
|    |    |    + Information URL (*)
|    |    |    |    + Name (0)
|    |    |    |    + URL (1)
|    |    |    |    + Description (0)
```

SPASE-Based Searching

Searches across VOs based on the SPASE Data Model can be done by several methods:

- Any of the VOs can harvest and store all SPASE Data Descriptions and provide a cross-disciplinary search through their search interface.

- Registries can collect SPASE Resource Types and support searches

  - SPASE Resource Types: Catalog
    Display Data
    Numerical Data
    Granule
    Instrument
    Observatory
    Person
    Registry
    Repository
    Service

- SPASE Query Language can be used to search generically across the VOs, - assuming use of or mapping of SPASE metadata

Virtual Magnetospheric Observatory

FEATURES

- VMO has basic keyword and time search

- Again based on SPASE data set descriptions

- Can search by major resource types

- Data quickly downloadable

http://vmo.igpp.ucla.edu/search/

Virtual Heliospheric Observatory

FEATURES

- **VHO emphasizes query building**

- **Query categories based on SPASE terminology**

- **Can save custom-built query**

- **Data granules available in XML and plot form**

http://vho.nasa.gov/index.php/data-query-mainmenu-26

or google "virtual heliospheric observatory" and look for "Data Search"

11

Virtual Space Physics Observatory

FEATURES

- **Contains all SPASE data set descriptions**

- **Uses SPASE terminology for search keywords**

- **Provides data access through "get data" links**

- **Queries easily modified**

- **Can ingest XML-based SPASE descriptions**

http://vspo.gsfc.nasa.gov/websearch/dispatcher

or google "VSPO"

Application Tools

Tools for working with SPASE metadata and the SPASE framework.

Validator
Determines compliance with a version of the SPASE data model.

XML Validate

Parser
Convert SPASE XML to internal structures

Parser

Editor
Web-based Editors

Web Editor

Standalone Editors

SPASE Assistant

Editors with Database Storage

Web+DB Editor

Generator
Create SPASE descriptions using external sources of information

Ruleset Description Generator

Harvester
Extracts information from SPASE resource descriptions (or registries)

SPASE Registry Server

SPASE Database Query

Wrapper
Converts or embeds SPASE metadata in other descriptions or forms (i.e., OAI)

Data Dictionary Lookup

SPASE-to-OAI mapping

Correlator
Divide an XML document into individual resource descriptions into a well organized file system

Correlator

Refcheck
Determine the validity of all references in a resource descriptions. Checks Resource IDs and URL

Refcheck

There are additional tools in development:

SPASE Query Language

Java-to-XML Binding Mechanism (JAXB)
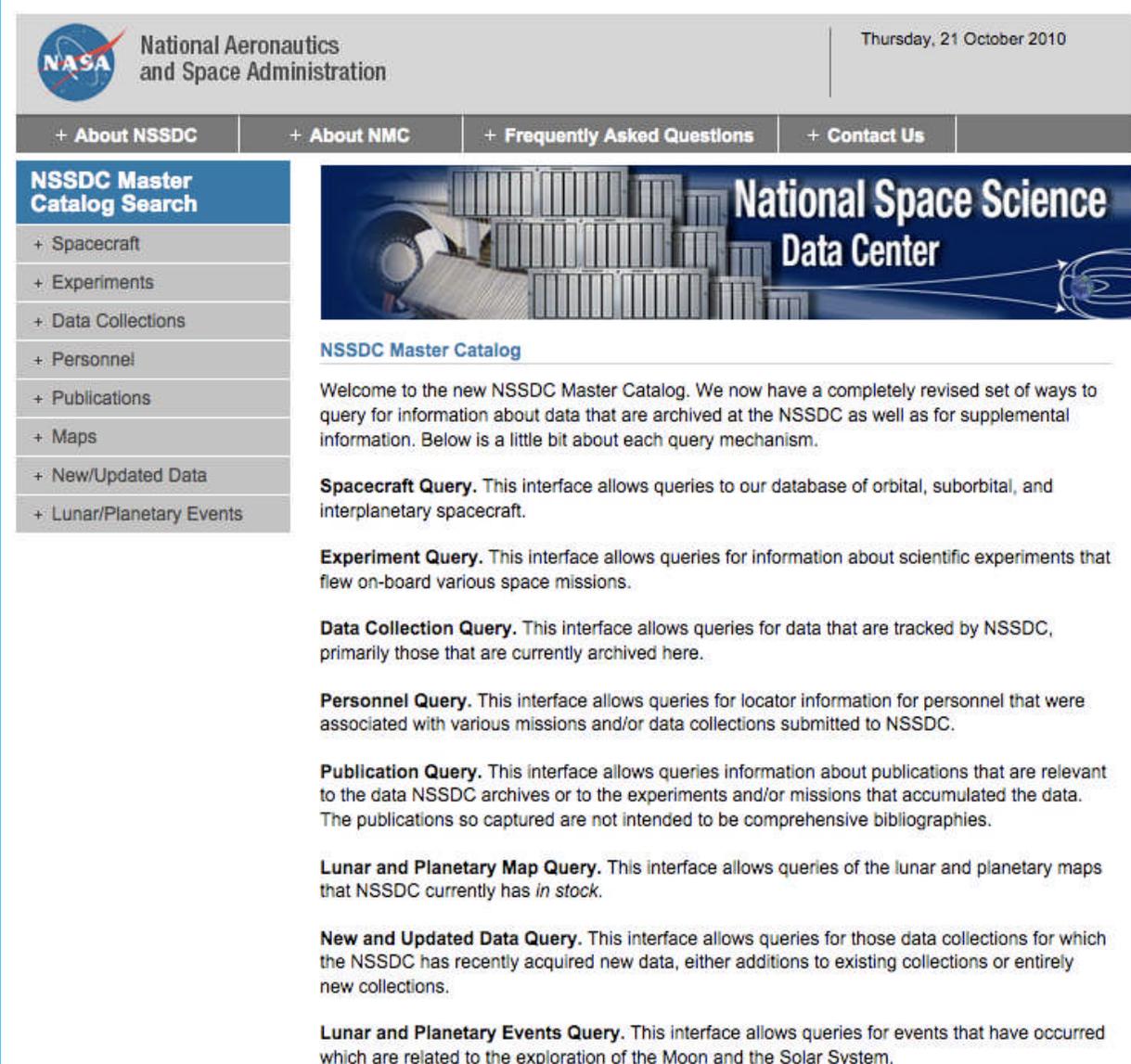
SPASE Guidelines Document

NSSDC Role in SPASE

- The National Space Science Data Center (NSSDC) leads the SPASE effort (biweekly teleconferences, periodic face-to-face meetings, presentations at appropriate conferences, etc.)

- Maintaining the "deep archive" for heliophysics and planetary data

- Providing SPASE data set descriptions for popular NSSDC-requested data sets

- Enabling the SPASE Registry of spacecraft, instrument, and personnel information from the *World Data Center for Satellite Information*

NSSDC – World Data Center

- The NSSDC is also the World Data Center for Satellite Information (WDC SI)

- The NSSDC/WDC SI **Master Catalog** holds information about rocket, satellite, and space probe launches; satellite orbit elements and ephemerides; descriptions of spacecraft and experiments

- WDC SI also publishes the **SPACEWARN** Bulletin which summarizes spacecraft launches monthly (http://nssdc.gsfc.nasa.gov/spacewarn/)

NSSDC – Master Catalog

- **Provides searchable information on:**
  - satellites
  - experiments
  - associated personnel
  - etc.

- **Satellite ephemerides are often available**

- **Information is exportable for SPASE Registries**

http://nssdc.gsfc.nasa.gov/nmc/

Trusted Digital Repositories

… there is a demand for a standard against which Repositories of digital information may be audited and on which an international accreditation and certification process may be audited.

- Developing a standard with criteria that a repository archive must meet to be an ISO Trusted Digital Repository

- Being developed as a Consultative Committee for Space Data Systems (CCSDS) and ultimately an ISO standard

- Also includes requirements for groups providing audit and certification of TDRs

Audit and Certification of Trustworthy Digital Repositories (CCSDS/ISO Metrics Document)

- Provides a set of metrics to test an archives' compliance to OAIS archiving principles

- Archives can perform self-audits with this standard to identify their strong areas and areas where they may need improvement

- Certification requires an audit by an approved outside party with archives auditing experience

12/7/2010

Summary

- **The Heliophysics Data Environment is historically diverse and widespread, but is being unified**

- **SPASE provides a unified metadata approach**

- **Testing of the utility of the system in finding and acquiring data is underway (Fall AGU)**

- **Creation of SPASE data descriptions is key to progress.**

- **Utility of SPASE tools can greatly influence success**

- **NSSDC/World Data Center for Satellite Information provides valuable information for SPASE descriptions**

- **Standard for auditing and certifying data repositories is in progress**

BACKUP

Search Scenario

Assume the user is interested in finding a wide variety of data relevant to a study of the major solar storms from Halloween Oct. 31, 2003.

- Time span of search - 2003-10-30 00:00:00 to 2003-11-01 23:59:59

How easily can this presently be done?

Approaches:

- Search archive by archive --> impractical!

- Use Virtual Observatories one-by-one (sample searches follow)

- Use a SPASE-related search across VOs

VO Searches
Virtual Solar Observatory

FEATURES

- **Predated SPASE Data Model**

- **Assumes reasonable knowledge of solar physics observables**

- **Retrieval of data can be ordered**

Search Strategy: Use Time and Observable Search to find Halloween, 2003 Data-   Time search alone returned 4878 items, having both narrow and broad time range (just in Halloween time frame or a time span of years)

http://sdac.virtualsolar.org/cgi-bin/search

VO Searches
Virtual Ionosphere, Thermosphere, Mesosphere Observatory

FEATURES

- **Simple interface for searching**

- **Connects to Virtual Solar Terrestrial Observatory**

- **Aim for a higher level of search capability - a "discovery" system**

- **Intending to enable image-based searching**

Search Strategy: A special Halloween Storm 2003 search is provided under coverage but additional parameter must be specified - including all Near-Earth parameters yielded 26 "products" and 252 "results"

http://mizar.jhuapl.edu/sras/frameset.jsp

SPASE Registry

The SPASE Registry is a service for providing spacecraft, instrument, and personnel information for use in data set descriptions. Much of the information comes from the NSSDC Master Catalog and Personnel database.

Present

- Next release will support:
    - Forms interface
    - Ability to display results in raw XML or in XHTML form
    - Specification of SPASE version
    - Choice of NSSDC or SMWG IDs for resource IDs
- Completed all but last item which is in progress
- Release expected to be late March 2009

Future

- Configuration of git repository on NSSDC computer
- Setting up git repository to synchronize with VMO repository
- Creation of software to update NIMS database from git repository contents for observatories, instruments, people
- Creation of software to update git repository contents based on updates to NIMS database

Overview – Past Year

Data Model
- Official release: 2.1.0
- Current draft: 2.2.0
  - Add "Hardcopy" as a format
  - Add "Operating Span" to Instrument and Observatory
  - Add coordinate systems for solar physics (HCC, HCR, HPC, HPR)
  - Update definitions
  - More changes under consideration (i.e., S3_Bucket)

Software
- Release of SPASE toolkit

Web site
- New design for website (ready to implement)

Issues

- How well does SPASE function as an *interlingua* among the Virtual Observatories and data archives?
- How effective is SPASE in describing data sets for data finding and usage?
- How much should SPASE be "inside" vs. "outside" the observatories, etc. to be effective?
- To what level of detail should data descriptions be created in order to fulfill the objectives of SPASE and the Space Physics data environment?
- What else is needed in SPASE for the non-NASA virtual observatories environment?
- Should SPASE data descriptions be centrally stored or distributed?
- What interfaces are most effective for SPASE searches?
- Should SPASE be expanded for usage in closely-related disciplines such as Planetary Science?

ABSTRACT

Session H2: <u>The ICSU World Data System: a Disciplined or Un-disciplined Intitiative?</u>

Title of H2-3: The Heliophysics Data Environment, Virtual Observatories, NSSDC, and SPASE

Heliophysics (the study of the Sun and its effects on the Solar System, especially the Earth) has an interesting data environment in that the data are often to be found in relatively small data sets widely scattered in archives around the world. Within the last decade there have been more concentrated efforts to organize the data access methods and create a Heliophysics Data and Model Consortium (HDMC). To provide data search and access capability a number of Virtual Observatories (VO's) have been established both via funding from the U.S. National Aeronautics and Space Administration (NASA) and through other funding agencies in the U.S. and worldwide. At least 15 systems can be labeled as Heliophysics Virtual Observatories, 9 of them funded by NASA. Other parts of this data environment include Resident Archives, and the final, or "deep" archive at the National Space Science Data Center (NSSDC). The problem is that different data search and access approaches are used by all of these elements of the HDMC and a search for data relevant to a particular research question can involve consulting with multiple VO's – needing to learn a different approach for finding and acquiring data for each.

The Space Physics Archive Search and Extract (SPASE) project is intended to provide a common data model for Heliophysics data and therefore a common set of metadata for searches of the VO's and other data environment elements. The SPASE Data Model has been developed through the common efforts of the HDMC representatives over a number of years. We currently have released Version 2.1 of the Data Model. The advantages and disadvantages of the Data Model will be discussed along with the plans for the future. Recent changes requested by new members of the SPASE community indicate some of the directions for further development.

Search Mechanism Across Virtual Observatories

- Many Virtual Observatories in Heliophysics, both new and old - often quite different from each other
- Many important Heliophysics Data Centers are not yet directly connected with a Virtual Observatory

To make data "findable" in the Heliophysics Data Environment a common metadata language and descriptions of data sets in the metadata language are important (SPASE)

- The Space Physics Archive Search and Extract (SPASE) project is an international collaboration begun through the NASA-funded Heliophysics Virtual Observatories initiative

Activities of SPASE

Tools
- SPASE Editor
- SPASE Validator
- SPASE Referential checker.
- SPASE Collator.

Engineering/Design
- SPASE data model. (dictionary, documents, tools)

Services
- SMWG (Core entity descriptions)
- SPASE services (reference implementation)
    - resolver, downloader, render, status,
    - jetty server, explorer

Content
- SPASE publication list
- Website

THE WEBSITE AND THE SPASE DATA MODEL

Version 1.4 of the Data Model is imminent and will be "frozen" for community usage.

http://spase.gsfc.nasa.gov

World Data Centers Relationship

- Solar-Terrestrial World Data Centers are important to the Solar and Space Physics Data Environment

- Not often a relationship with a Virtual Observatory

- Many other important Heliophysics Data Centers are also not likely to be directly connected with a Virtual Observatory

To make data "findable" in the Heliophysics Data Environment a common metadata language and descriptions of data sets in the metadata language are important (SPASE)

- The Space Physics Archive Search and Extract (SPASE) project is led by the National Space Science Data Center which is also the World Data Center for Satellite Information

Current Status

Development of the SPASE Data Model continues in biweekly teleconferences and occasional face-to-face meetings

- Version 1.2.1 is frozen for use in data description
- New developments are incorporated in Version 1.3
- Data descriptions using the SPASE Data Model are being made in a number of places
- Issues resulting from data description problems are usually addressed and resolved within a few weeks
- NSSDC provides SPASE support from its budget
- NSSDC provides spacecraft, instrument, and personnel information for registries
- Tools for assisting in the use of SPASE continue to be developed
- The SPASE core group is considering the level of detail necessary to support data descriptions

Roles for SPASE

- Current focus is the SPASE data model.
- Demonstration technologies.

- Keep scope or expand?