2.03. The Virtual Solar Observatory: The Best Laid Schema

J.B. Gurman, G. Dimitoglou
NASA Goddard Space Flight Center

F. Hill, S. Wampler
National Solar Observatory

R.S. Bogart, K. Tian
Stanford University

P. Martens, A. Davy
Montana State University
How many slides can I show?

• Why a VSO?
• How do we go about building a VSO?
  • Mostly boring history and “programmatics”
• What would a VSO look like?
• VSO design concepts original, strawman
• An introduction to some of the technical glue that will hold the VSO together
  • XML
  • XML schemas
  • SOAP
• The road ahead
There are only two conceivable reasons for data “centers:”
- Unique services/expertise
- Services too expensive to provide on a distributed basis

In the 1980’s and early 1990’s, both of these applied to the SDAC
- Neither is true any longer

A distributed model, with multiple instances of data services when possible, is more robust
- A distributed model also allows colocation of data and expertise

We’re all really Groucho Marxist John Lennonists, right? To wit, we wouldn’t want to belong to any club that would have us as members, and we all want a revolution in the cost and accessibility of data services.
VSO: How? I Who’s paying for it?

- A bit of really boring history
  - First attempts organized c. 1995 by K. Reardon and L. Sanchez Duarte as the “Whole Sun Catalog”
    - unfunded by European sources
  - BoF session at 2000 SPD Lake Tahoe, proposal effort led by F. Hill NSO
    - unfunded by NSF
  - Parallel effort led by R. Bogart Stanford
    - unfunded by NASA LWS DATM now TR&T
  - Proposed as leading to the “withering away” of the SDAC
    - funded by NASA SEC senior review 2001
VSO: How? II
Mechanics

• Three unsolicited proposals  Stanford, NSO, Montana State, 2002 May 2003 April
  • Reviewed by VSO Steering Committee
  • Biweekly telecons, face to face meetings every 6 months
  • Produced “strawman” presented at last December’s AGU
• New contract  NSO lead, others subcontracting, 2003 May
**VSO: How? III People**

**Steering Committee**
- Todd Hoeksema  NASA HQ
- Rob Bentley, chair  MSSL, UK; EGSO
- Sam Freeland  LMSAL
- Steve Walton  CSUN
- Dominic Zarro  L 3 EER/GSFC

**NASA**
- Chuck Holmes  MO&DA Program Manager
- Joe Gurman  GSFC; de facto project scientist

**VSO Team**
- Rick Bogart, Karen Tian  Stanford
- Frank Hill, Steve Wampler  NSO
- Piet Martens, Alisdair Davey  MSU
- Joe Gurman, George Dimitoglou  GSFC

**And you**
- Your comments at AAS, AGU, SPD sessions and BoF’s; e.g. this meeting
- Any input/any time
- Community testing/ adoption or not
VSO: What?

- It should be distributed
- It should let us search for and access data from multiple missions/observatories/wavelengths without intimate knowledge of the data organization
- It should provide access to analysis software, instrument descriptions, &c. that enable use of the data for research
- It should be easy to add new data sets
- Given the funding profile for this effort it has to attempt to draw a “small box” around a small set of attributes that are useful for doing science
What a VSO won’t do at least at first

- Data Mining
- Remote processing
  - can be a “research opportunity,” e.g. through CoSEC funded by LWS TR&T: cf. poster 3.09
- Grid computing
  - as opposed to data grids: NSO and GSFC are members of EGSO
- Access control ever The VSO will be in the business of giving data away
VSO: Original Design Concept

- Three parts:
  - distributed archives
  - metadata “broker” facility
  - Web based front end

- Could have different implementations:
  - peer to peer
  - grid
  - &c.

- Would be low cost
  - < 1.2M over four years proposed

Either we were really smart (same conceptual model adopted by NVO, PDS) ....or maybe there was one obvious way to do this.
1. Access through a browser or an API
2. “Small box” uses registry of XML data service schema to construct appropriate queries for each relevant data service
3. API or browser can refine queries
4. Final data transfer is direct to requestor no middleman

See poster 3.11 Davey et al.
What is XML?

• XML is a markup language, as is HTML, but:
  • HTML describes how to display data, while
  • XML describes data in a natural format for metadata
• In addition,
  • HTML consists of predefined tags, but
  • in XML, tags have to be defined
• XML uses XML schema XSD to describe data
Some XML

• This is some of the XML used to produce this presentation in Apple Keynote, before it was exported to Powerpoint.
What is an XML schema?

- An XML Schema defines the allowed components of an XML document.

- An XML Schema defines:
  - elements that can appear in a document
  - attributes that can appear in a document
  - which elements are child elements
  - the order of child elements
  - the number of child elements
  - whether an element is empty or can include text
  - data types for elements and attributes
  - default and fixed values for elements and attributes
An XML schema

```xml
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
   version="1.0">
   <xs:annotation>
      <xs:documentation xml:lang="en">
         VSO data service registry
      </xs:documentation>
   </xs:annotation>

   <xs:complexType name="Observable">
      <xs:sequence>
         <xs:element name="Description" type="xs:string"/>
      </xs:sequence>
      <xs:attribute name="Poe" type="xs:string"/>
   </xs:complexType>

   <xs:complexType name="Dopplergram">
      <xs:complexContent>
         <xs:extension base="Observable">
            <xs:sequence>
               <xs:element name="Polarization" type="Polarization"/>
            </xs:sequence>
         </xs:extension>
      </xs:complexContent>
   </xs:complexType>
</xs:schema>
```

- Part of the VSO data service registry
- Describes searchable data types, elements, and their attributes
- Goes on quite a bit farther e.g. time coverage, dataset name, provider, &c.
- Types can be added as new data types appear
A schema instance
single data service

<?xml version="1.0"?>
<DataSources xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
           xsi:noNamespaceSchemaLocation="registry.xsd">
  <DataProvider>
    <Name>Stanford</Name>
    <Organization>Stanford University</Organization>
    <Facility>Instrument</Facility>
    <Address>Stanford, California</Address>
    <Contact>R. Bogart</Contact>
    <Dataset Name="MDI">
      <Dopplergram Poe="http://vso.stanford.edu/cgi-bin/hmm4.pl?page=0;go=Next;observable=Dopplergram">
        <Description>Dopplergram</Description>
      </Dopplergram>
      <Magnetogram Poe="http://vso.stanford.edu/cgi-bin/hmm4.pl?page=0;go=Next;observable=Magnetogram">
        <Description>Magnetogram</Description>
      </Magnetogram>
      <Continuum Poe="http://vso.stanford.edu/cgi-bin/hmm4.pl?page=0;go=Next;observable=Continuum">
        <Description>Continuum</Description>
      </Continuum>
      <ContinuumFiltergram Poe="http://vso.stanford.edu/cgi-bin/hmm4.pl?page=0;go=Next;observable=ContinuumFiltergram">
        <Description>ContinuumFiltergram</Description>
      </ContinuumFiltergram>
      <LineDepth Poe="http://vso.stanford.edu/cgi-bin/hmm4.pl?page=0;go=Next;observable=LineDepth">
        <Description>LineDepth</Description>
      </LineDepth>
      <TimeCoverage>
        <Start>1996-01-01</Start>
        <End>2003-02-01</End>
      </TimeCoverage>
      <OperationStatus>On-line</OperationStatus>
      <Distribution>HTTP</Distribution>
    </Dataset>
  </DataProvider>
</DataSources>

- Part of the VSO registry
- Describes SOHO MDI dataset in terms of observables
- Includes Points of Entry
- What happens if I point a browser at a PoE?

Search MDI lev1.5 data by observables

- Observables
  - Continuum
  - ContinuumFiltergram
  - Dopplergram
  - Magnetogram
  - LineDepth

Next
What is SOAP?

- Simple Object Access Protocol
  - Designed for communication between applications
  - A format for sending messages
  - Platform, language independent
  - Machine, rather than human interface oriented
  - Based on XML, and therefore extensible
  - As far as VSO is concerned:
    - One layer in a Web services protocol stack
    - Nicely supported in perl
  - See VSO poster 3.12 Tian et al.
How do we get o here to there?

• Study
  • Completed 2002 November; approved by Steering Committee 2003 January

• Contract

• Prototype implementation, FY03/FY04
  • Four sites multiple data services
  • Test usefulness/usability with community

• IFF the prototype proves useful to the community
  • Continue limited development
  • Add “research opportunities” separately funded
    • Distributed processing e.g. CoSEC
    • Connections with other efforts EGSO, LWS DS
  • Extended maintenance phase
    • Add nodes
    • Support old, new nodes
What will it take to join the VSO?

• You have a network accessible archive
  • VSO should be able to help small data services get online
• Your observations can be described with a few parameters understandable by most solar physicists: name, date and time, frequency, &c.
• In current model, will require registering and running a SOAP server
  • Data service as Web service
VSO Resources

• VSO homepage
  • http://virtualsolar.org/

• Strawman proposal
  • http://virtualsolar.org/docs/VSO_strawman_20021125.pdf

• Sample XML schemas:
  • http://virtualsolar.org/docs/schemas/