EGSO in need for a global schema

A. Csillaghy[*1], J. Aboudarham[2], E. Antonucci[3], R. D. Bentley[4], A. Finkelstein[4], L. Ciminiera[5], J. B. Gurman[6], F. Hill[7], I. Scholl[8], D. Pike[9], V. Zharkova[10]
[1]Univ. of Applied Sciences (Switzerland); [2]Observatoire de Paris-Meudon (France);
[3]Osservatorio Astronomico di Torino (Italy), [4]Univ. College London (UK);
[5]Politecnico di Torino (Italy); [6]NASA Goddard Space Flight Ctr. (USA);
[7]National Solar Observatory (USA); [8]Institut d'Astrophysique Spatiale (France);
[9]Rutherford Appleton Lab. (UK); [10]Univ. of Bradford (UK)

ABSTRACT

The European Grid of Solar Observations (EGSO) is a project to develop a virtual observatory for the solar physics community. Like all such projects, a vital component is a schema that adequately describes the data in the distributed data sets. Here, we discuss the schema in general terms, and present a draft example of a portion of a possible XML schema.

Keywords: virtual observatories, XML schema, distributed archives, solar data description

1. INTRODUCTION

Although the Internet has made it much easier to share data, the rapidly increasing volume and complexity of solar data necessitate a seed change in the way solar data are handled. The task of identifying data sets of interest, then locating and retrieving them, remains a continuing difficulty. The data are heterogeneous and widely distributed, without any means to tie them together. In addition, there is no systematic way to identify observations associated with a particular feature of type of event.

The European Grid of Solar Observations (EGSO) is designed to confront these issues. It will allow a user to identify solar observations covering a given time interval and pointing, or a type of feature; it will locate the selected observation and then return them after any necessary pre-processing. To achieve its objectives the project will develop unified observing catalogues and the tools to search them, and will federate data archives to simplify the recovery of the data.

2. THE EGSO

The EGSO is a project to provide a coordinated community-wide resource for accessing and analyzing solar observations. It will be capable of sharing resources coming from all types of data providers, while ensuring scalability and compatibility among all datasets. Users will be given the access to a range of solar and heliospheric data archives. They will be able to browse among heterogeneous and distributed solar data sets. In essence, the EGSO will create the fabric of a virtual observatory.

One of the major hurdles facing solar physicists is the task of locating and combining the available data from numerous individual sites (Sanchez Duarte et al. 1997). This problem becomes only more critical as new space- and ground-based instrumentation substantially increases the volumes of data that are produced. It is clear that a federation of solar data archives around the world would be of great advantage in addressing a broad range of scientific questions. EGSO will organize the selection, process and retrieval of these distributed and heterogeneous solar data sets using global observing catalogues, more appropriately called a global schema in ORDBMS terms, for

[*] csillag@fh-aargau.ch; phone 011 056 462 44 11; fax 011 056 462 44 15; http://www.fh-aargau.ch; FH-Aargau Technik, Klusterzelgstrasse, CH-5210 Windisch, Switzerland.

space and ground-based observations (citation). Moreover, EGSO will also apply image recognition techniques to identify particular solar features (e.g. filaments, coronal streamers, etc.), creating new catalogs of events and phenomena with which to select solar data of interest.

The EGSO is a European-Commission-funded Information Society Technologies (IST) grid test bed project that began in March 2002 and will last three years. EGSO will also work closely with groups funded under the NASA Virtual Solar Observatory (VSO) initiative and the Living With A Star (LWS) program. In addition EGSO will interact with many similar projects such as the European Virtual Observatory (VO), the US National Virtual Observatory (NVO), and Astrogrid. These projects are primarily driven by the current and planned nighttime large-scale survey projects, but there are many cross-cutting issues such as system architecture, metadata standards, security, usage reporting, data rights, and generalized data descriptors (i.e. VOTable). The major differences between nighttime and solar Virtual Observatories lie in the object catalog, coordinate systems, and emphasis on the temporal domain. There are also similarities between the EGSO and earth and planetary science projects such as Environmental Information Systems.

The EGSO is currently in the concept definition phase. Figs. 1 and 2 provide block diagrams of the current architecture concept for the query resolver and locating and processing the data.

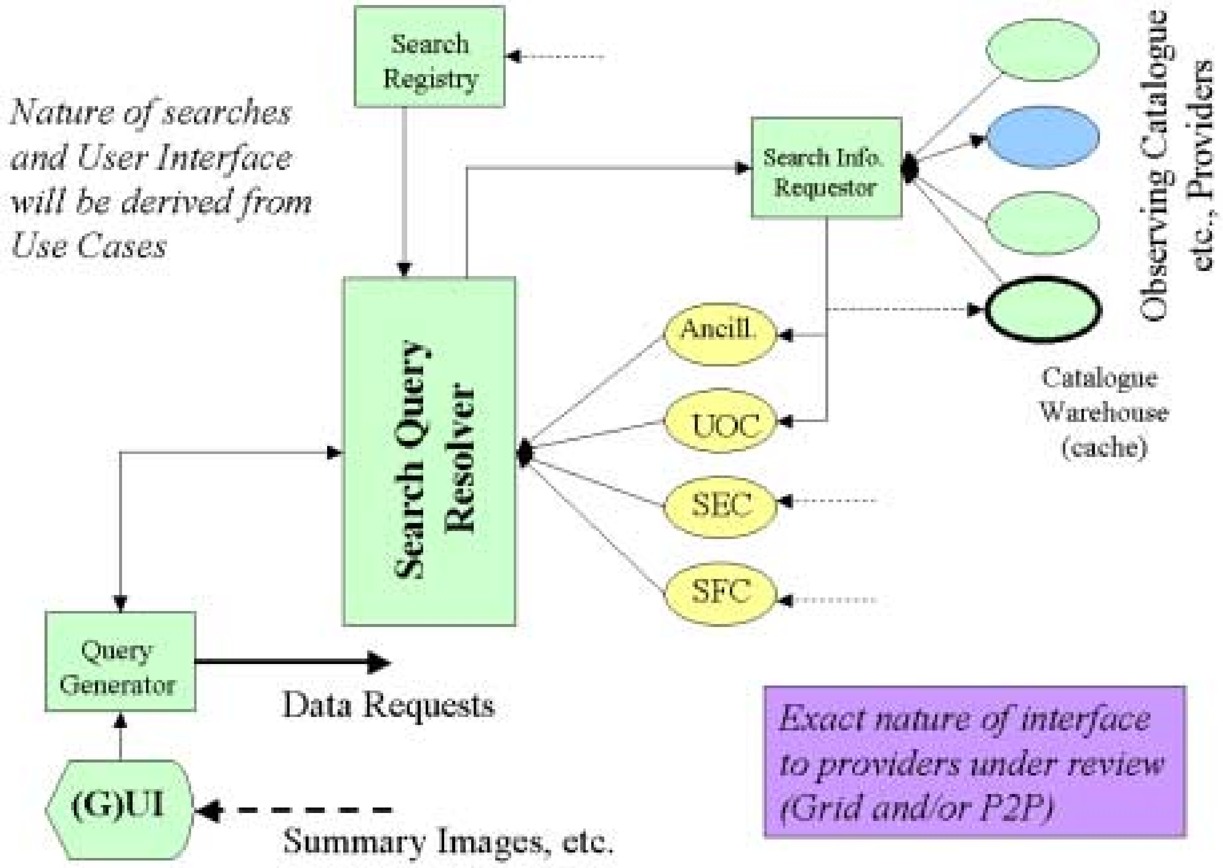

Search Registry
Nature of searches and User Interface will be derived from Use Cases
Search Info. Requestor
Observing Catalogue etc., Providers
Search Query Resolver
Ancil.
UOC
SEC
SFC
Catalogue Warehouse (cache)
Query Generator
Data Requests
(G)UI
Summary Images, etc.
Exact nature of interface to providers under review (Grid and/or P2P)

Fig. 1: The EGSO conceptual block diagram for resolving a query from a user via a user interface which could be either graphical or command-line driven [(G)UI]. The query is processed via a resolver that has inputs from a Solar Feature Catalog (SFC), a Solar Event Catalog (SEC), a Universal Observing Catalog (UOC), and other ancillary catalogs. Once these inputs are collated and resolved, the information is passed to a Search Information Requestor, which sends queries to the warehouse of catalogs of the distributed data nodes.

The nature of the searches and the requirements for the user interface are being derived from the development of several specific use cases generated by solar physicists. These use cases will yield a relatively small number of general characteristics that will set the functional requirements for the interface and for optimizing the searches. For instance, a set of commonly used typical searches will be derived and provided to users to reduce the gradient of the learning curve. The user interface will likely have two modes – a graphical mode for data exploration, and a command-line mode to allow direct integration of EGSO functionality into user community software such as SolarSoft (Freeland and Handy 1998). The exact nature of the interface between the EGSO and the data nodes is under study. The leading contenders are currently Grid technology and peer-to-peer (P2P) approaches.

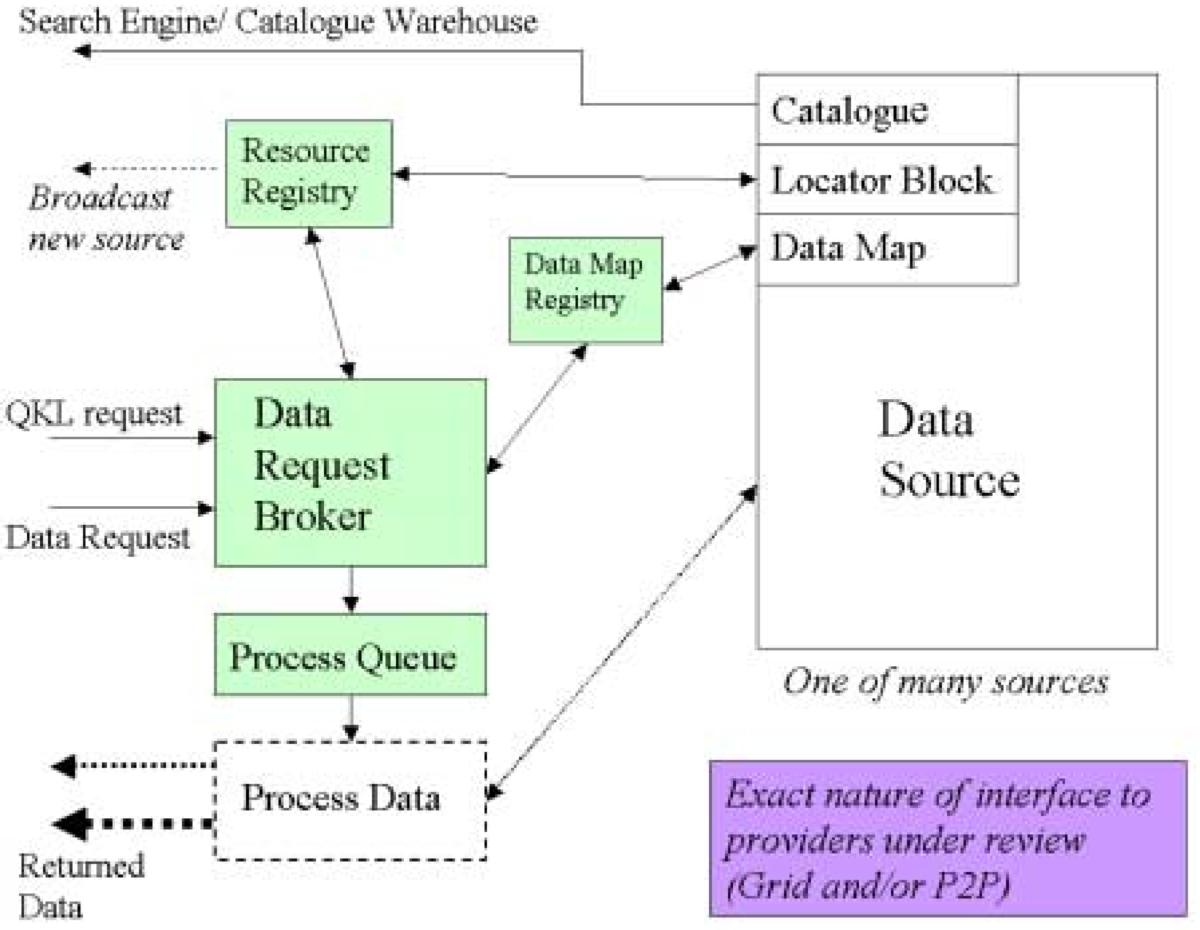


Fig. 2: The EGSO conceptual block diagram for locating and processing the data from a user query. The Data Request Broker is a key element that must be aware of the details of the data source catalogs, metadata, formats, and holdings.

It is quite clear from Figs. 1 and 2 that there must be detailed specifications of the catalogs, metadata, formats, and holdings for each data source, and that these specifications should be constructed in an easily extensible way to allow the addition of new data sources. This is the motivation for designing a global schema for solar data.

3. DATA MODELING REQUIREMENTS

One approach to developing a virtual observatory is to first focus on modeling the data, and then adding the functionality and processing capabilities later. This approach borrows from object-relational systems, with the advantage that the data is not randomly distributed. Solar data possess an inherent organization so, by first developing a correct data model, we can build a system that organizes itself along these classes. The data model determines from a data point of view how the system is organized. Therefore, this is a data-centric view instead of the more common view of function-centric.

In solar physics, space and ground-based observations are both important. The observations are made at many different wavelengths, and can be either of the full disk or smaller fields-of-view. The different wavelengths originate from different levels in the solar atmosphere and combining information from them allows the user to build up a three-dimensional picture of changes in structure, motion of material, the sites of energy release, etc. Satellite-based observations are often made at wavelengths that do not penetrate the Earth's atmosphere. Ground-based observations are important in that they compliment the satellite-based observations. Most optical and radio observations are only made from the ground – indeed some require facilities that are not currently possible in space. Finally, there are helioseismic observations, from both ground and space, which have a somewhat different approach to describing the acoustic spectra and the products derived from them.

Satellites are normally operated under the umbrella of large organizations such as space agencies, and the instruments they carry are often built by international collaborations. As a consequence, from the outset data are handled in a more systematic and open manner. The data are reformatted by the instrument or observatory and are stored in archives – these are often at mission level, with copies at one or more sites. The files have various formats, including FITS, and range from single images to extended intervals (an hour or orbit). Access to these data varies. Some archives have web-based interfaces, but the effectiveness of these varies wildly. Usually some form of observing catalogue (lists of observing modes) is available, but there are currently no standards for the format of these.

The ground-based observatories involved are both large and small, and are located throughout the world, scattered over many time zones. Since any observatory only observes for a fraction of the day due of the rotation of the Earth, and because they are often affected by weather, good coverage at radio and optical wavelengths often means dealing with a number of observatories. The data are generally available as FITS files, often of single images. Normally there is a single copy of the data, managed by observatory. There is limited access to the data through the Internet, and since cataloging is generally less systematic than for satellite observations, determining what observations were actually made at a given observatory can be a problem.

4. EGSO AND GLOBAL SCHEMA

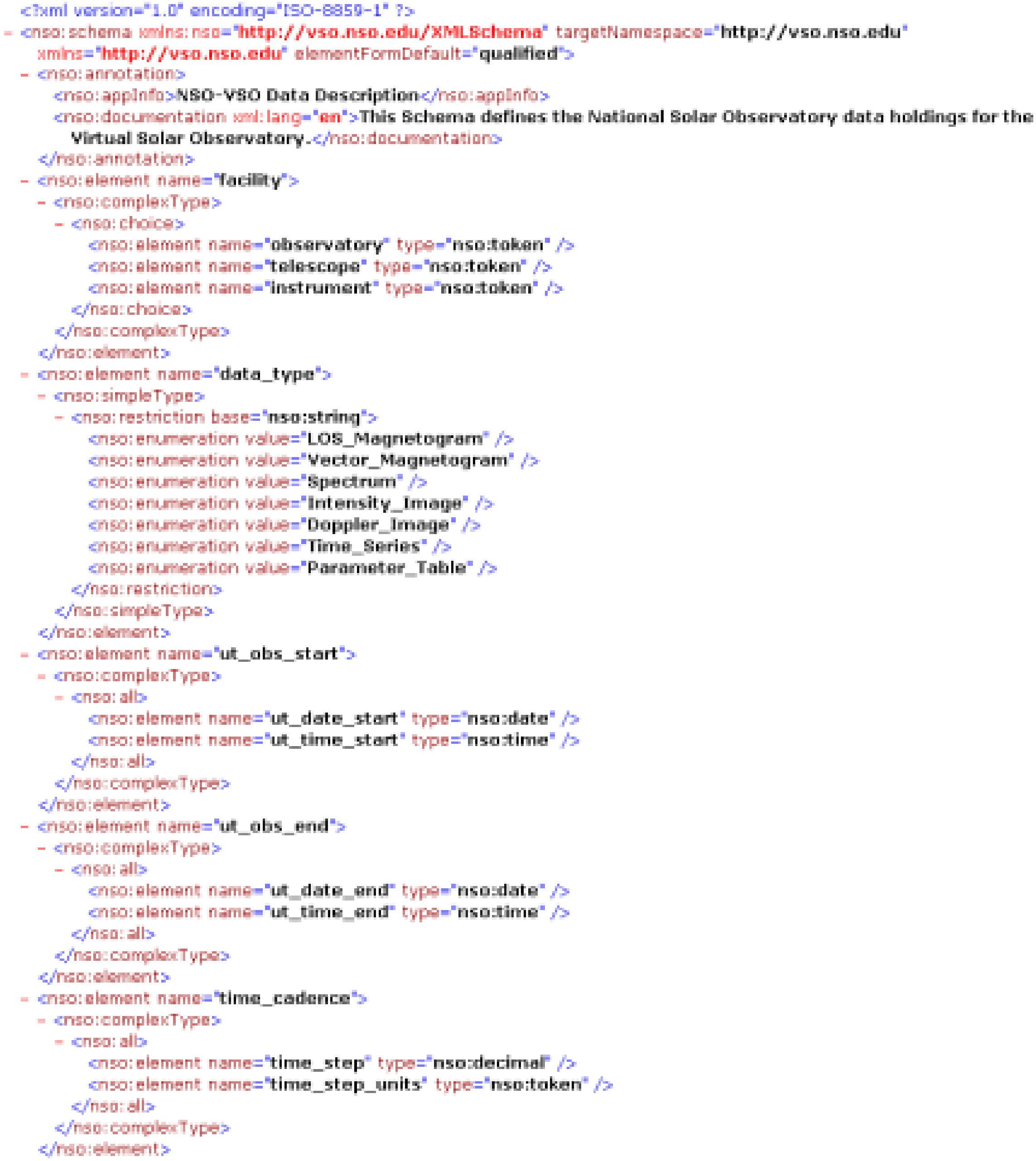
The EGSO must thus deal with an extremely heterogeneous, large, and unevenly collated data set. It thus requires a global schema that would model all of these observations in a generic way. The EGSO system is a typical case where the data can be put at the center. There are hundreds of different data sources, but the concepts (entities, relationships, and attributes) behind the observations are much more limited. What can be described with a relational schema can also be stored in an RDBMS.

The data modeling allows us to distinguish between the relational part and the semi-structured part of the system. Note that the global schema is not really derived from the actual data. Instead, it ties together the different existing (local) catalogues and schemas using wrappers. Wrappers are configurable software components that are responsible for this mapping. In addition, the development of a global schema provides a framework for the user interface structure by specifying the search categories, ranges, and formats. In fact, it is feasible to derive a schema through use cases and the consequent demands a user would make of an interface in various research scenarios.

The choice of language in which to write the EGSO schema is fairly obvious. XML (eXtensible Markup Language) is a markup language for creating unique tag sets, a data exchange format, a metalanguage for creating information-rich documents, and a way to exchange middleware messages. With an XML–based schema, the EGSO could exploit a rapidly growing set of standards and tools, such as the Simple Object Access Protocol (SOAP), to construct some of the core pieces of the system.

5. SCHEMA EXAMPLE: DESCRIBING THE NSO HOLDINGS

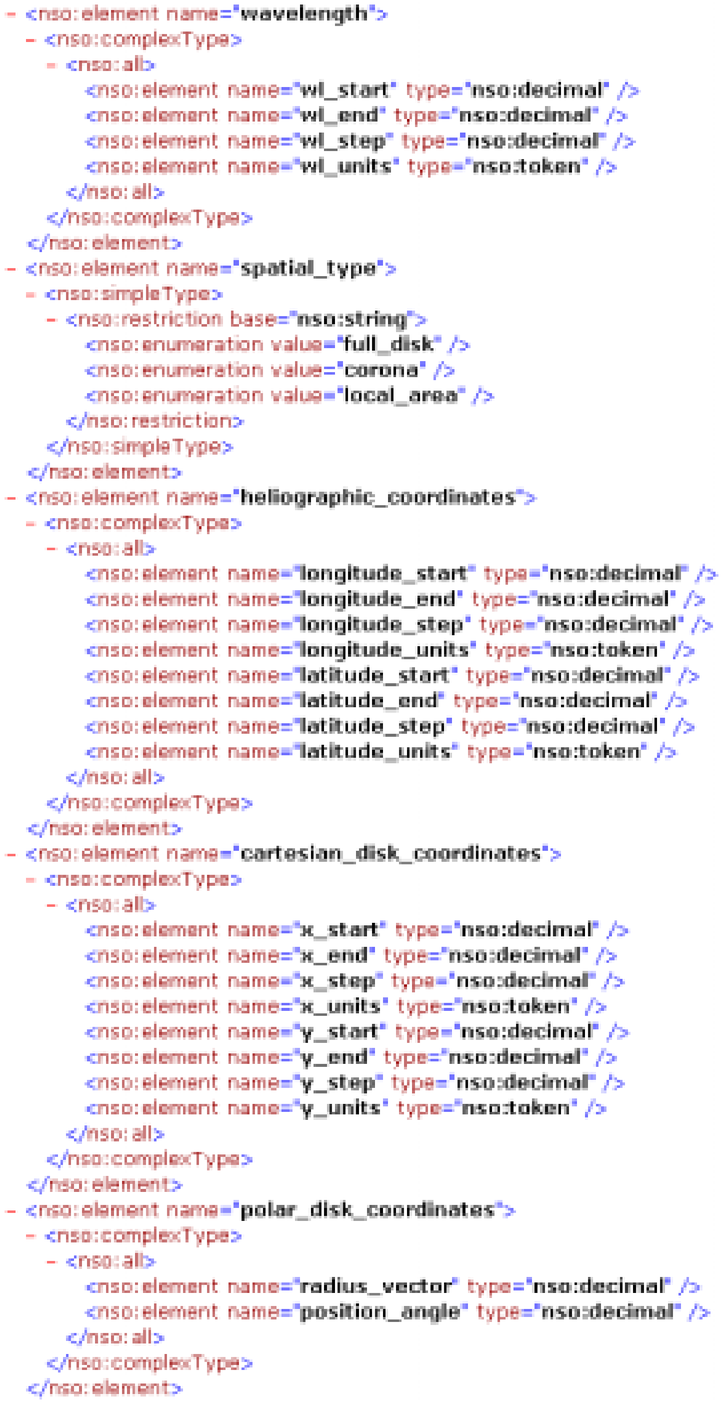
In addition to the EGSO global schema, each data source must construct a local schema that both conforms to the global specifications, and adequately describes the local data node structure and contents. To illustrate one concept of a portion of the data model, we present a draft XML schema developed for the National Solar Observatory (NSO) holdings (Hill et al. 2000) in the context of the VSO. These holdings consist of full-disk line-of-sight magnetograms in two wavelengths, He 1083-nm intensity images, Ca K and Hα images, spectra from the Fourier Transform Spectrometer, and helioseismic data from the Global Oscillation Network Group (GONG) project. In addition, the NSO data set will soon hold the output from SOLIS, which will add vector magnetograms to the set. The schema is shown in Figs. 3, 4, and 5.



```
  <?xml version="1.0" encoding="ISO-8859-1" ?>
- <nso:schema xmlns:nso="http://vso.nso.edu/XMLSchema" targetNamespace="http://vso.nso.edu"
    xmlns="http://vso.nso.edu" elementFormDefault="qualified">
  - <nso:annotation>
      <nso:appInfo>NSO-VSO Data Description</nso:appInfo>
      <nso:documentation xml:lang="en">This Schema defines the National Solar Observatory data holdings for the
        Virtual Solar Observatory.</nso:documentation>
    </nso:annotation>
  - <nso:element name="facility">
    - <nso:complexType>
      - <nso:choice>
          <nso:element name="observatory" type="nso:token" />
          <nso:element name="telescope" type="nso:token" />
          <nso:element name="instrument" type="nso:token" />
        </nso:choice>
      </nso:complexType>
    </nso:element>
  - <nso:element name="data_type">
    - <nso:simpleType>
      - <nso:restriction base="nso:string">
          <nso:enumeration value="LOS_Magnetogram" />
          <nso:enumeration value="Vector_Magnetogram" />
          <nso:enumeration value="Spectrum" />
          <nso:enumeration value="Intensity_Image" />
          <nso:enumeration value="Doppler_Image" />
          <nso:enumeration value="Time_Series" />
          <nso:enumeration value="Parameter_Table" />
        </nso:restriction>
      </nso:simpleType>
    </nso:element>
  - <nso:element name="ut_obs_start">
    - <nso:complexType>
      - <nso:all>
          <nso:element name="ut_date_start" type="nso:date" />
          <nso:element name="ut_time_start" type="nso:time" />
        </nso:all>
      </nso:complexType>
    </nso:element>
  - <nso:element name="ut_obs_end">
    - <nso:complexType>
      - <nso:all>
          <nso:element name="ut_date_end" type="nso:date" />
          <nso:element name="ut_time_end" type="nso:time" />
        </nso:all>
      </nso:complexType>
    </nso:element>
  - <nso:element name="time_cadence">
    - <nso:complexType>
      - <nso:all>
          <nso:element name="time_step" type="nso:decimal" />
          <nso:element name="time_step_units" type="nso:token" />
        </nso:all>
      </nso:complexType>
    </nso:element>
```

Fig. 3: Part 1 of a draft schema describing the NSO data holdings

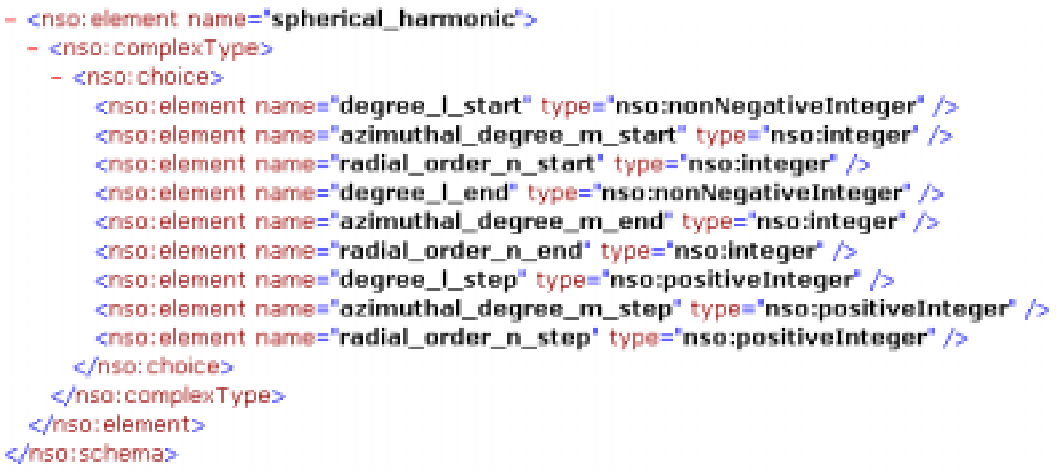
Fig. 3 shows the first part of the schema. This part contains a preamble specifying the URL of the schema name space, and an annotation with a short description of the schema. Next come schema elements describing the source and type of the data, with an enumerated list of the allowable data types in the NSO data set. The last three elements specify the temporal characteristics of the data.



```
- <nso:element name="wavelength">
  - <nso:complexType>
    - <nso:all>
        <nso:element name="wl_start" type="nso:decimal" />
        <nso:element name="wl_end" type="nso:decimal" />
        <nso:element name="wl_step" type="nso:decimal" />
        <nso:element name="wl_units" type="nso:token" />
      </nso:all>
    </nso:complexType>
  </nso:element>
- <nso:element name="spatial_type">
  - <nso:simpleType>
    - <nso:restriction base="nso:string">
        <nso:enumeration value="full_disk" />
        <nso:enumeration value="corona" />
        <nso:enumeration value="local_area" />
      </nso:restriction>
    </nso:simpleType>
  </nso:element>
- <nso:element name="heliographic_coordinates">
  - <nso:complexType>
    - <nso:all>
        <nso:element name="longitude_start" type="nso:decimal" />
        <nso:element name="longitude_end" type="nso:decimal" />
        <nso:element name="longitude_step" type="nso:decimal" />
        <nso:element name="longitude_units" type="nso:token" />
        <nso:element name="latitude_start" type="nso:decimal" />
        <nso:element name="latitude_end" type="nso:decimal" />
        <nso:element name="latitude_step" type="nso:decimal" />
        <nso:element name="latitude_units" type="nso:token" />
      </nso:all>
    </nso:complexType>
  </nso:element>
- <nso:element name="cartesian_disk_coordinates">
  - <nso:complexType>
    - <nso:all>
        <nso:element name="x_start" type="nso:decimal" />
        <nso:element name="x_end" type="nso:decimal" />
        <nso:element name="x_step" type="nso:decimal" />
        <nso:element name="x_units" type="nso:token" />
        <nso:element name="y_start" type="nso:decimal" />
        <nso:element name="y_end" type="nso:decimal" />
        <nso:element name="y_step" type="nso:decimal" />
        <nso:element name="y_units" type="nso:token" />
      </nso:all>
    </nso:complexType>
  </nso:element>
- <nso:element name="polar_disk_coordinates">
  - <nso:complexType>
    - <nso:all>
        <nso:element name="radius_vector" type="nso:decimal" />
        <nso:element name="position_angle" type="nso:decimal" />
      </nso:all>
    </nso:complexType>
  </nso:element>
```

Fig. 4: Part 2 of the draft schema.

Fig. 4 shows the second part of the schema. Here the wavelength characteristics are specified. The general element set of “start”, “end”, “step” and “units” are a common feature of many of the elements in solar schema. Next the spatial characteristics of the data are described. This section is generally the most complex of solar schema due to the variety of coordinate systems in use.



```
- <nso:element name="spherical_harmonic">
  - <nso:complexType>
    - <nso:choice>
        <nso:element name="degree_l_start" type="nso:nonNegativeInteger" />
        <nso:element name="azimuthal_degree_m_start" type="nso:integer" />
        <nso:element name="radial_order_n_start" type="nso:integer" />
        <nso:element name="degree_l_end" type="nso:nonNegativeInteger" />
        <nso:element name="azimuthal_degree_m_end" type="nso:integer" />
        <nso:element name="radial_order_n_end" type="nso:integer" />
        <nso:element name="degree_l_step" type="nso:positiveInteger" />
        <nso:element name="azimuthal_degree_m_step" type="nso:positiveInteger" />
        <nso:element name="radial_order_n_step" type="nso:positiveInteger" />
      </nso:choice>
    </nso:complexType>
  </nso:element>
</nso:schema>
```

Fig. 5: Part 3 of the draft NSO schema.

Fig. 5 shows the last part of the draft NSO schema. Here one major component of the helioseismology data set is described, namely the global acoustic spectra of the sun.

This draft example is incomplete. Items that are missing include more helioseismology products, and a more detailed specification of the actual contents and search parameter limits of the NSO data set. However, this example gives a flavor of the type of elements that must be included in the EGSO set of schema. In reality, the global schema should be developed first and the data set schema then constructed to conform to the global specifications.

6. CONCLUSION

The EGSO will provide a major advance in the research tool kit of solar physics. By allowing a user to search several distributed data sets without having to learn the details of each individual data set, the EGSO will facilitate large-scale correlative statistical research on solar phenomena. In order to accomplish this, the EGSO needs a comprehensive, general, and compact description of solar data sets. Developing an XML schema, such as the one for the National Solar Observatory, provides an adequate solution to this need.

ACKNOWLEDGEMENTS

The EGSO project is funded under the Information Society Technologies (IST) thematic program of the European Commission's Fifth Framework Program (FP5) contract No. IST-2001-32409. NSO is operated by the Association of Universities for Research in Astronomy, under a cooperative agreement with the National Science Foundation, for the benefit of the astronomical community.

REFERENCES

1. Freeland, S.L. and Handy B.N., “SolarSoft”, *Solar Physics*, **182**, p. 497, 1998.
2. Hill, F., Erdwurm, W. Branston, D., and McGraw, R. “The National Solar Observatory Digital Library – A resource for space weather studies”, *J. Atmosph. Solar-Terrestrial Phys.*, **62**, p. 1257, 2000.
3. Sanchez Duarte, L., Fleck, B. and Bentley, R., “The Whole Sun Catalogue”, *Proceedings of 1st Advances in Solar Physics Euroconference, ASP Conf. Ser.* **119**, p. 382, 1997.