University of Huddersfield Repository

Beech, Guy

The Benefits of Using XML Technologies in Astronomical Data Retrieval and Interpretation

Original Citation

Beech, Guy (2016) The Benefits of Using XML Technologies in Astronomical Data Retrieval and Interpretation. In: Proceedings of the 2016 conference on Big Data from Space (BiDS'16). Publications Office of the European Union, Joint Research Centre, pp. 268-271. ISBN 978-92-79-56980-7

This version is available at http://eprints.hud.ac.uk/27913/

The University Repository is a digital collection of the research output of the University, available on Open Access. Copyright and Moral Rights for the items on this site are retained by the individual author and/or other copyright owners. Users may access full items free of charge; copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational or not-for-profit purposes without prior permission or charge, provided:

- The authors, title and full bibliographic details is credited in any copy;
- A hyperlink and/or URL is included for the original metadata page; and
- The content is not changed in any way.

For more information, including our policy and submission procedure, please contact the Repository Team at: E.mailbox@hud.ac.uk.

http://eprints.hud.ac.uk/

THE BENEFITS OF USING XML TECHNOLOGIES IN ASTRONOMICAL DATA RETRIEVAL AND INTERPRETATION

Guy Beech
School of Computing and Engineering, University of Huddersfield, HD1 3DH, UK

ABSTRACT

This paper describes a solution found during recent research that could provide improvements in the efficiency, reliability and cost of retrieving stored astronomical data. This solution uses XML Technologies in showing that when querying a variety of astronomical data sources a standardised data structure can be output into an XML query results Document. This paper shows the astronomical XMLSchema that has been partially developed in conjunction with simple custom supporting system software. It also discusses briefly possible future implications.

Index Terms - XML Schema, standard, data processing

1. INTRODUCTION

The current retrieval processes of existing astronomical records have a number of difficulties that are shown in the table below:

Multiplicity of formats
Transformation problems due to incompatible formats
Archaic problems with vintage data such as limited storage space, maximum string length and other restrictions of technology of their time
Need for digitisation
The increasing rate of astronomical data capture both by institutions and amateurs
The growing different types of astronomical data collectors [1]

Table 1: Retrieval Difficulties

If these could be solved it would result in more reliable and cost effective acquisition of useful and valuable data.

Although there are many institutions and individuals using good standards of astronomical data capture, there is a lack of any widely adopted comprehensive standards of astronomical data recording that are currently in use. Due to this variety of storage systems data retrieval is often expensive and incomplete. [2]

The purpose of the research carried out was to:

- Identify a precise way forward to achieving improved efficiency of data retrieval, through gaining understanding of current literature, developing a new retrieval process and working towards an astronomical schema to be adopted as standard. [3]

The literature review of this research confirmed that the great majority of all records are now held in digital format [4] in electronic databases, in a wide variety of data structures and that there has also been in recent years an extremely rapid increase in the rate at which new data is being saved, which is a trend that is predicted to carry on into the future. [5]

The literature review also indicated that there is no widely used, comprehensive standard of data recording used in astronomy. Schemas produced so far, including those of the Virtual Observatory, have been limited in scope in that they have been designed to cover designated sections of astronomy and not astronomy as a whole. Details of the evaluation of existing schemas are to be found in my thesis. [3] A single schema that is capable of providing a standard data structure across all of the branches of astronomy will take time to develop. Therefore care was taken during the research that in the initial stages of schema development the design of the structure enables later additions to it. The following Hypothesis was tested:

- That an extensible schema can be developed to provide a common data retrieval structure for astronomical data, it can be designed in such a way that it will extend to cover all areas of astronomical data and this schema can be shown to work within a software system which ensures that the queried data retrieved is subjected to schema validation.

This study was developed from a consideration of biochemistry information retrieval schemas in a paper by Marco Mesiti and colleagues which discusses XML solutions for the problems of the representation, integration and management of heterogeneous biological data, involving biological data types represented by a number of XML languages and schemas. These models discussed gave good guidance as they were developed to handle solutions to the problems of large amounts of disparate data over many disciplines, [6] not unlike the situation with Astronomical data.

2. THE LOGICAL MODEL

The extensible schema was developed for use as a controlling mechanism which only permits the use of astronomical data once it had been retrieved from databases and saved into an approved structure via schema validation. An XMLSchema was considered suitable for this for the reasons that Diagram 1 illustrates:

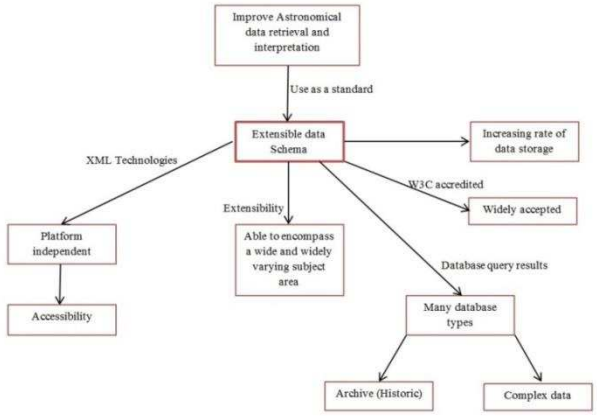

Diagram 1: Advantages of an XML schema

A custom software application developed as part of this research would be used to test the performance of the schema.

3. IMPLEMENTATION OF THE PHYSICAL MODEL

The structure of the astronomy XMLSchema was dictated by the known primary branches of Astronomy and sub branches that were needed for the astronomical data that was used during this research.

The testing that was carried out retrieved data of visible and infra-red type observations and so the XMLSchema was developed with the following structure in those sub-branches:

Astronomy (root)
Astrophysics
Astrobiology
Astrochemistry
Historical Astronomy
Planetary Science
Observational Astronomy
Visible
Infrared
Radio
Microwave
Gravity wave
Shortwave
Neutrino
Submillimetre

Table 2: Astronomical Schema high level nodes

All the branches of Astronomy other than 'Observational Astronomy' can be developed at some time subsequently as can all the types of 'Observational Astronomy' observations, other than 'Visible' and 'Infrared' types which were fully developed to the following detailed structure as shown in Figure 1 below:

Visible
- device
  - manufacturer
  - model
- location
  - gridref
  - address
- target
  - targetname
  - targetposition
- targetephemeris
  - ephemeris
- datetime
  - thedate
  - thetime
- weather
  - weatherdesc
- observer
  - forename
  - surname
  - contact
- visibleobservationdata
  - iscolour
  - magnitude
  - description
  - fileurlvisible
  - filenamevisible
  - thedatavisible
  - thedatavisibleblob (blob data)
- images
  - filename
  - fileurl
  - imagedata (blob data)
- comments
  - thecomments

Infrared
- device
  - manufacturer
  - model
- location
  - gridref
  - address
- target
  - targetname
  - targetposition
- targetephemeris
  - ephemeris
- datetime
  - thedate
  - thetime
- weather
  - weatherdesc
- observer
  - forename
  - surname
  - contact
- irobservationdata
  - thetime
  - theintensity
  - thewavelength
  - fileurlir
  - filenameir
  - thedatair
  - thedatairblob (blob data)
- images
  - filename
  - fileurl
  - imagedata (blob data)
- comments
  - thecomments

Figure 1: Detail of Visible and Infra-red nodes

The inclusion of BLOB data (proposed as Base 64 encoding) storage nodes is particularly valuable as it allows for streams of data to be added to the xml that it is not possible to add as 'string' data.

The software developed to demonstrate the XMLSchema in a practical scenario was given the name Retrieval of Astronomical Data (ROAD). The structure of this software system is shown below in Diagram 2:

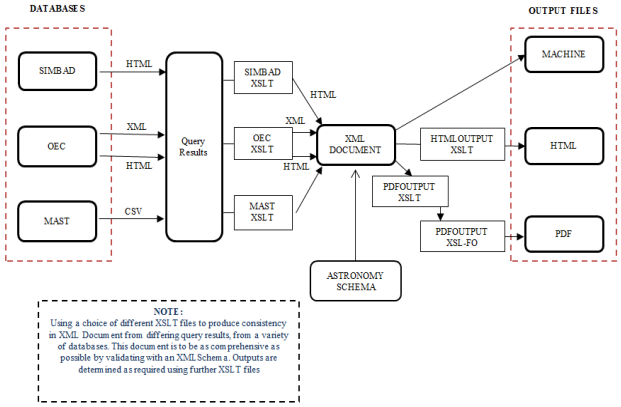

Diagram 2: ROAD system structure

The ROAD system currently has very basic, limited capabilities which are designed to just prove the concepts. A properly capable user friendly system has yet to be developed. It is able to query up to three separate databases, (those of OEC, MAST and SIMBAD) combining the returned data into a query results object which is then used to create an XMLDocument of standard structure containing data, as defined by the XSLT transformation files for each database. The user of the ROAD system can choose up to three output file types of PDF, HTML and MACHINE (which is basically XML for the purpose of machine to machine communication). The astronomy XMLSchema scans the XMLDocument and only permits the production of output files if the XMLDocument is of the approved structure.

The diagram above shows there are three different data types that were retrieved from the databases of HTML, CSV and XML which comprised four different formats in total as the HTML query responses were in different formats. The data was originally captured by both Optical and Infra-red telescopes. The table below shows those database data difficulties resolved by the ROAD system:

Multiplicity of formats	✓ use of XSLT
Transformation problems due to incompatible formats	✓ use of XSLT
Archaic problems with vintage data such as limited storage space, maximum string length and other restrictions of technology of their time	✓ schema nodes
Need for digitisation	Not resolved
The increasing rate of astronomical data capture both by institutions and amateurs	Not resolved
The growing different types of astronomical data collectors [1]	✓ schema nodes

Table 4: Retrieval Difficulties Resolved

4. RESULTS

Five test runs were made of the ROAD system of differing combinations of the databases being queried and different outputs requested. The results are shown in Table 5 below.

Run	Choice of Sources	Choice of Outputs	Actual Outputs
1	MAST SIMBAD OEC_XML OEC	HTML PDF Machine	resultsXmlDocument_15 03 2015 19-08-48.xml htmlOutput_15 03 2015 19-08-50.html machineOutput_15 03 2015 19-08-50.xml pdfOutput_15 03 2015 19-08-50.fo run1output.pdf (using Attic)
2	MAST SIMBAD OEC_XML	HTML	resultsXmlDocument_15 03 2015 19-13-28.xml htmlOutput_15 03 2015 19-13-30.html
3	SIMBAD OEC_XMLOEC	PDF	resultsXmlDocument_15 03 2015 19-17-05.xml pdfOutput_15 03 2015 19-17-06.fo run3output.pdf (using Attic)
4	MAST, OEC_XML	Machine	resultsXmlDocument_15 03 2015 19-24-50.xml machineOutput_15 03 2015 19-24-51.xml
5	MAST SIMBAD OEC_XML OEC.	HTML PDF Machine	resultsXmlDocument_15 03 2015 19-35-19.xml Schema error here

Table.5: ROAD Run Results

It can be seen from the table that runs 1 to 4 produced outputs of types chosen by the user within the available range of PDF, HTML and Machine (XML).

Note that in Run 5 the XMLDocument creation was deliberately made to be produced in an invalid format (by making changes to the XSLT files) and therefore no Output files were produced. Shown here is the log output for that test run which also gives information of the reason that the XMLDocument was found to be invalid:

```
==========================================
ROADS beginning data retrieval 15/03/2015 19:35:19
==========================================
Querying for data...
Querying for Mast
Querying for OEC
```

Querying for OEC_XML
Querying for Simbad
Query results returned
The element 'infrared' has invalid child element 'visualobservationdata'. List of possible elements expected: 'irobservationdata'.
resultsXmlDoc did not validate

XMLDocument created: C://roadLogs//resultsXmlDocument_15 03 2015 19-35-19.xml

*** Error *** No file outputs due to XMLDocument schema validation error
==========
ROADS finished data retrieval 15/03/2015 19:35:21
==========

Figure 2: Log output of Run 5

5. CONCLUSIONS and FUTURE DEVELOPMENTS

This extensible XMLSchema is capable of further development without compromising the structure created so far. This is due to the fact that the root node is that of 'astronomy' with all the different areas of astronomy as child nodes. More child nodes can be added or existing ones further developed in the future as they are needed.

All the test runs correctly produced one XMLDocument each containing combined data from the datasource queries that were chosen by the user. Output files were also produced from each run. The first four test runs correctly produced the output files requested by the user. The final fifth run did produce a schema error as expected and correctly, in accordance with expectations, did not create any user output files.

This research resolved several of the data retrieval difficulties by:

- Creating a partial astronomy XMLSchema
- Implementing the proof of concept ROAD software and then carrying out the test runs that retrieved data into XMLDocuments validated by the XMLSchema prior to any output of data to users.

It is this XMLSchema validation process that provides the control on the XMLDocument structure to ensure a standard structure. It is this standardization of structure that can reduce the loss of metadata (by ensuring that it is included) and enable easier retrieval and use. Additionally it is possible that building up a generally available library of XMLDocuments from completed database queries could be of benefit for accessing data more easily and completely. Also, building up an available library of XSLT files to be used for querying astronomical datasources and generating user outputs could be an increasingly valuable asset for the creation of XMLDocuments containing astronomical data. Also MACHINE type outputs could be a way of automating requests directly to telescopes and other types of detecting instruments to collect new data, if deficiencies in existing data were found.

Future Research

It is apparent that future research is required for the full development of the XMLSchema and a resource of XSLT and XMLDocument libraries. Such research would require skills of both a technical and sociological nature. On the technical side the XMLSchema needs to be extended across the full spectrum of astronomical disciplines and more XSLT files and data queries need to be developed. On the sociological side, this solution to the data retrieval problem needs to be attractive to the astronomical research and industry sectors. Achieving the W3C standard for the XML Schema would be a big step forward in achieving widespread acceptance.

6. REFERENCES

[1] VAO. (2015, May 1). History of the Virtual Astronomical Observatory. Retrieved from VAO: http://virtualobservatory.org/whatis/history.aspx

[2] NASA. (2014, October 21). XDF: The Extensible Data Format Based on XML Concepts. Retrieved from NASA: http://nssdc.gsfc.nasa.gov/nssdc_news/june01/xdf.html

[3] Beech, G. (2015, October). An Investigation of the Benefit of XML Technologies in Astronomical Data Interpretation. Huddersfield University, West Yorkshire, UK.

[4] Lopez, M. H. (2014, November 14). *The World's* Technological Capacity to Store, Communicate, and Compute Information. Retrieved from Science Magazine: http://www.sciencemag.org/content/332/6025/60

[5] Wall, M. (2014, November 17). Astronomy Overload. Retrieved from Space.com: http://www.space.com/9308-astronomy-overload-scientists-shifting-stargazing-data-mining.html

[6] Mesiti, M. (2011). XML-Based Approaches for the Integration of Heterogeneous Bio-Molecular Data. In H. H. Trimm, Recent Advances in Biochemistry (pp. 206 -). CRC Press.