JHOVE2: Next-Generation Architecture for Format-Aware Characterization

SGML Module Specification

Version 2.0.1
Issued 2011.01.05
Status Final
Copyright © 2010 by The Regents of the University of California, ITHAKA, and The Board of Trustees of Leland Stanford Junior University. All rights reserved.

This work is licensed under the Creative Commons Attribution-Share Alike 3.0 United States License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/3.0/us/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

JHOVE2 Project Team

California Digital Library
Stephen Abrams
Patricia Cruse
John Kunze
Isaac Rabinovitch
Marisa Strong
Perry Willett

Portico
John Meyer
Sheila Morrissey

Stanford University
Richard Anderson
Tom Cramer
Hannah Frost

Library of Congress
Martha Anderson
Justin Littman

With help from
Walter Henry
Nancy Hoebelheinrich
Keith Johnson
Evan Owens
Preface

JHOVE2 is a Java framework and application for next-generation format-aware characterization. Characterization is the process of examining a formatted digital object and automatically extracting or deriving representation information about that object that is indicative of its significant nature and useful for purposes of classification, analysis, and use. For more information, visit http://jhove.org.

This document covers the specification, design, implementation, and configuration of the JHOVE2 SGML format module.

Acknowledgments

The JHOVE2 project is funded by the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP).

Version History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0.0</td>
<td>2010.09.10</td>
<td>Draft</td>
</tr>
</tbody>
</table>
# Contents

Preface 3  
Acknowledgments 3  
Version History 3  
Contents 4  
1 Introduction 5  
2 Identification 5  
3 References 6  
4 Terminology and conventions 7  
5 Validity 8  
6 Format profiles 8  
7 Reportable Properties 8  
8 Configuration 22  
   Configuring the OpenSp Parser 22  
   Configuring ONSGMLS options 23  
   Configuring OSGMLNorm Options 23  
9 Implementation Notes 23  
   Wrapping OpenSP 23  
   The SGML Catalog 24  
   Using ANTLR to Parse ESI output 24  
Appendix A: OpenSP ONSGMLS Options 25  
ONSGMLS 25  
   SYNOPSIS 25  
   DESCRIPTION 25  
   OPTIONS 25  
Appendix B: OpenSP OSGMLNORM Options 33  
OSGMLNORM 33  
   SYNOPSIS 33  
   DESCRIPTION 33  
   OPTIONS 34  
Appendix C: OpenSP Output Format (ESI) 35  
Onsgmls Output Format 35
1 Introduction

This is the specification for the JHOVE2 Standard Generalized Markup Language (SGML) module. SGML is an ISO-standard (ISO 8879:1986 SGML) technology for defining generalized markup languages for documents.

NOTE: The JHOVE2 SGML module comprises a wrapper around the OpenSP SGML parser, originally developed by James Clark. This is a free and open-source library, written in C++. The Open SP parser is not included in the JHOVE2 distribution. Please see the JHOVE2 User’s Guide for instructions on how to download and install OpenSp prior to invoking the SGML module.

Important information for users of the JHOVE2 SGML module

The authoritative specification for SGML is unambiguous. A format specification is considered unambiguous if there is broad community consensus regarding the intention of all normative requirements of the format’s authoritative specification; otherwise it is considered ambiguous, and areas of potential ambiguity will be documented below.

Validation of SGML instances by this module is comprehensive. Module validation is considered comprehensive if all normative requirements defined by its format’s authoritative specification are validated; otherwise it is considered selective and non-validated features will be documented below.

2 Identification

<table>
<thead>
<tr>
<th>Primary format or format family</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Canonical format name</strong></td>
</tr>
<tr>
<td><strong>Alias format name(s)</strong></td>
</tr>
<tr>
<td><strong>Canonical format identifier</strong></td>
</tr>
<tr>
<td><strong>Alias format Identifier(s)</strong></td>
</tr>
<tr>
<td>[PUID] x-fmt/195</td>
</tr>
<tr>
<td>[MIME] application/sgml</td>
</tr>
<tr>
<td>[RFC] RFC 1874</td>
</tr>
<tr>
<td>[ISO] ISO 8879:1986</td>
</tr>
<tr>
<td>[PUID] fmt/96 (Hypertext Markup Language)</td>
</tr>
<tr>
<td>[PUID] fmt/97 (Hypertext Markup Language V2.0)</td>
</tr>
<tr>
<td>[PUID] fmt/98 (Hypertext Markup Language V3.2)</td>
</tr>
<tr>
<td>[PUID] fmt/99 (Hypertext Markup Language V4.0)</td>
</tr>
<tr>
<td>[PUID] fmt/100 (Hypertext Markup Language V4.01)</td>
</tr>
</tbody>
</table>
### Format profile or specify format with a family

<table>
<thead>
<tr>
<th>Canonical format name</th>
<th>N/A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alias format name(s)</td>
<td></td>
</tr>
<tr>
<td>Canonical format identifier</td>
<td></td>
</tr>
<tr>
<td>Alias format Identifier(s)</td>
<td></td>
</tr>
</tbody>
</table>

### SgmlModule module

<table>
<thead>
<tr>
<th>JHOVE2 module name</th>
<th>SgmlModule</th>
</tr>
</thead>
<tbody>
<tr>
<td>JHOVE2 module class</td>
<td>org.jhove2.module.format.sgml.SgmlModule</td>
</tr>
</tbody>
</table>

### 3 References

For the purposes of the JHOVE2 SGML module the following are considered authoritative:


http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387


4 Terminology and conventions

DTD

Document Type Definition. A set of mark-up declarations that formally defines an SGML document type. All valid SGML document conform to the DTD specified via either the System or Public identifiers in the document’s DOCTYPE declaration.

Catalog

A formatted file that enables mapping from System and Public identifiers in an SGML document to the actual location of the DTD. See http://www.w3.org/TR/REC-html40/sgml/intro.html for a sample catalog file for HTML DTDs.

OpenSP

OpenSP is an open-source C++ tool, based on James Clark’s SP SGML parser, which parses and validates SGML files. Its onsgmls utility produces output in ESIS format, detailing the key SGML-related properties of the document, and indicates if the document conforms to its DTD. Its osgmlnorm utility uses the same codebase to parse the file, and will extract the DOCTYPE statement as well. Both utilities produce .err files in the same format, with error messages, some of which include a message level code.

ESIS

ISO 8879 Element Structure Information Set. The set of information that is acted upon by implementations of structure-controlled applications is called the “element structure information set” (ESIS). As even the characters used for start and end tag indicators can vary from DTD to DTD (i.e., they need not be “angle brackets”), the ESIS provides a standard way of comparing the element structure of a document as seen by various SGML applications.

Please see “Appendix C: OpenSP Output Format (ESIS)” for detailed information about OpenSP representation of ESIS.

DSL

Domain-Specific Language. A programming language or specification language dedicated to a particular problem domain, a particular problem representation technique, and/or a particular solution technique. The ESIS and message output from OpenSP comprise DSLs.

ANTLR

ANTLR, ANother Tool for Language Recognition, is a JAVA language tool that provides a framework for constructing recognizers, interpreters, compilers, and translators from
grammatical descriptions containing actions (http://www.antlr.org/). The SGML module uses ANTLR grammars to parse the output and error files created by OpenSP, in order to extract the feature information about each SGML file.

5  Validity

An SGML document is valid if the document’s Document Type Definition (DTD) conforms to the SGML specification, and the document itself conforms to the DTD.

If the SGML module is unable to locate the DTD (either directly from information in the DOCTYPE statement in the SGML document, or indirectly via a catalog file), then the SGML document will be deemed invalid.

If any procedural errors occur in the processing of the SGML document (for example, an IO Exception, or some error detected by the parser), then validity of the SGML document will be deemed to be Undetermined.

6  Format profiles

There are no profiles of the SGML format as such. However, SGML is perhaps best understood as a meta-format. Any single SGML-conformant Document Type Definition (DTD) specifies a class of documents. That class itself can be considered to comprise a format.

So, for example, in its default configuration, JHOVE2 “recognizes” any file identified to it by DROID as HTML, HTML 2.0, HTML 3.2, HTML4.0, or HTML 4.01 as belonging to the SGML format family

7  Reportable Properties

SgmlModule Class

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/Validator/isValid">http://jhove2.org/terms/property/org/jhove2/module/format/Validator/isValid</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>org.jhove2.module.format.Validator.Validity</td>
</tr>
<tr>
<td>Description</td>
<td>Validation status.</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Property</td>
<td>Identifier</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>ReportableName Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/core/reportable/AbstractReportable/ReportableName">http://jhove2.org/terms/property/org/jhove2/core/reportable/AbstractReportable/ReportableName</a></td>
</tr>
<tr>
<td>Property</td>
<td>Identifier</td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td>Scope Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/Module/Scope">http://jhove2.org/terms/property/org/jhove2/module/Module/Scope</a></td>
</tr>
<tr>
<td>TimerInfo Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/Module/TimerInfo">http://jhove2.org/terms/property/org/jhove2/module/Module/TimerInfo</a></td>
</tr>
</tbody>
</table>
### ModuleNotFoundMessage Message
- **Identifier**: http://jhove2.org/terms/message/org/jhove2/module/format/BaseFormatModule/ModuleNotFoundMessage
- **Type**: org.jhove2.core.Message
- **Description**: Format Module Not Found Error Message
- **Reference**:

### ModuleNotFormatModuleMessage Message
- **Identifier**: http://jhove2.org/terms/message/org/jhove2/module/format/BaseFormatModule/ModuleNotFormatModuleMessage
- **Type**: org.jhove2.core.Message
- **Description**: Module returned is not Format Module Error Message
- **Reference**:

### isShouldFindDoctype Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlModule/isShouldFindDoctype
- **Type**: boolean
- **Description**: Parser setting: Run normalizer to construct DOCTYPE statement
- **Reference**:

### DocumentProperties Property
- **Type**: org.jhove2.module.format.sgml.SgmlDocumentProperties
- **Description**: SGML document properties
- **Reference**:

### SgmlParserErrorMessages Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlModule/SgmlParserErrorMessages
- **Type**: java.util.List<org.jhove2.core.Message>
- **Description**: Error-level messages returned by SGML parser
- **Reference**:

### SgmlParserWarningMessages Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlModule/SgmlParserWarningMessages
- **Type**: java.util.List<org.jhove2.core.Message>
- **Description**: Warning-level messages returned by SGML parser
- **Reference**:
### SgmlDocumentProperties Class

<table>
<thead>
<tr>
<th>Property</th>
<th>Identifier</th>
<th>Type</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ReportableName Property</strong></td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/core/reportable/AbstractReportable/ReportableName">http://jhove2.org/terms/property/org/jhove2/core/reportable/AbstractReportable/ReportableName</a></td>
<td>java.lang.String</td>
<td>Reportable name</td>
<td></td>
</tr>
<tr>
<td><strong>DocTypeFound Property</strong></td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/DocTypeFound">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/DocTypeFound</a></td>
<td>boolean</td>
<td>Was doctype statement determined</td>
<td></td>
</tr>
<tr>
<td><strong>isSgmlValid Property</strong></td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/isSgmlValid">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/isSgmlValid</a></td>
<td>boolean</td>
<td>SGML document conforms to its DTD</td>
<td></td>
</tr>
<tr>
<td><strong>PublicIdCount Property</strong></td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/PublicIdCount">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/PublicIdCount</a></td>
<td>int</td>
<td>Count of public identifiers associated with notations or external, text, or subdoc entities</td>
<td></td>
</tr>
<tr>
<td>Property Name</td>
<td>Identifier</td>
<td>Type</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>EntityFileNamesCount Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/EntityFileNamesCount</td>
<td>int</td>
<td>Count of system (file) identifiers associated with notations or external, text, or subdoc entities, generated by the entity manager from the specified external identifier and other information about the entity or notation</td>
<td></td>
</tr>
<tr>
<td>SysidsCount Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/SysidsCount</td>
<td>int</td>
<td>Count of system identifiers associated with notations or external, text, or subdoc entities</td>
<td></td>
</tr>
<tr>
<td>NotatDefCount Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/NotatDefCount</td>
<td>int</td>
<td>Count of notation names</td>
<td></td>
</tr>
<tr>
<td>ExtDataEntCount Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/ExtDataEntCount</td>
<td>int</td>
<td>Count of external data entity definitions</td>
<td></td>
</tr>
</tbody>
</table>

JHOVE2: Next-Generation Architecture for Format-Aware Characterization
### EntrefCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of external data entity references</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### subDocEntityDefCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of sub-document entity definitions</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### IntDataEntCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of internal data entity definitions</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### SubDocCommandCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of SGML subdocument entities</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### OmitCommandCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of omitted start-tag, end-tag, or attribute markup instances</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### ElementAttributeCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of attributes associated with elements</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>
### DataAttrCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of data attributes for a external entities</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### LinkAttrCount Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/LinkAttrCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of link attributes associated with elements</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### ElementCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of elements in SGML document</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### DataCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of element data content</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### IncludedSubElementsCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of included subelements</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### EmptyElementsCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of empty elements</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>
### CommentsCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of comments</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### SDataCount Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/SDataCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of internal SDATA entities</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### ProcessingInstructionsCount Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/ProcessingInstructionsCount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of processing instructions</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### AppInfoCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of APPINFO declared in DTD, and appearing in the text</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### PubIds Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.SortedSet&lt;java.lang.String&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Public identifiers associated with notations or external, text, or subdoc entities</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Property Name</td>
<td>Identifier</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>ExtEntFileNames Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/ExtEntFileNames">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocumentProperties/ExtEntFileNames</a></td>
</tr>
</tbody>
</table>
### EntRefNames Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument/Properties/EntRefNames
- **Type**: java.util.SortedSet<java.lang.String>
- **Description**: Names of external data entities referenced in SGML document
- **Reference**:

### SubDocEntDefNames Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument/Properties/SubDocEntDefNames
- **Type**: java.util.SortedSet<java.lang.String>
- **Description**: Names of defined sub-document entities
- **Reference**:

### SubDocCommandNames Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument/Properties/SubDocCommandNames
- **Type**: java.util.SortedSet<java.lang.String>
- **Description**: Names of defined sub-document entities appearing in SGML document
- **Reference**:

### ElementNames Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument/Properties/ElementNames
- **Type**: java.util.SortedSet<java.lang.String>
- **Description**: Unique names of elements appearing in SGML document
- **Reference**:

### SdataNames Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument/Properties/SdataNames
- **Type**: java.util.SortedSet<java.lang.String>
- **Description**: Unique SDATA entity names for SDATA entities appearing in SGML document
- **Reference**:

### InternalDataEntitytName2Value Property
- **Identifier**: http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument/Properties/InternalDataEntitytName2Value
- **Type**: java.util.Map<java.lang.String, java.lang.String>
- **Description**: Map from internal data entity name to entity value
- **Reference**:
### IntEnt2Type Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/IntEnt2Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.Map&lt;java.lang.String, java.lang.String&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Map from internal data entity name to entity type</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### InternalEntType2Count Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/InternalEntType2Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.Map&lt;java.lang.String, java.lang.Integer&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Map from internal data entity types to count of that type in document</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### ElemAttributeType2Count Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/ElemAttributeType2Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.Map&lt;java.lang.String, java.lang.Integer&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Map from element attribute types to count of that type in document</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### DataAttributeType2Count Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/DataAttributeType2Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.Map&lt;java.lang.String, java.lang.Integer&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Map from external entity data attribute types to count of that type in document</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### LinkAttributeType2Count Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/LinkAttributeType2Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.Map&lt;java.lang.String, java.lang.Integer&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Map from link attribute types to count of that type in document</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### ProcessingInstructions Property

<table>
<thead>
<tr>
<th>Identifier</th>
<th><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/ProcessingInstructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.List&lt;java.lang.String&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>Processing instructions in SGML document</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
<tr>
<td>Property</td>
<td>Identifier</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>PublicIdentifierFound Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/PublicIdentifierFound</td>
</tr>
</tbody>
</table>
### SystemIdentifierFound Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>boolean</td>
</tr>
<tr>
<td>Description</td>
<td>Indicates if system identifier for SGML document was found</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### SgmlParserMessages Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>java.util.List&lt;java.lang.String&gt;</td>
</tr>
<tr>
<td>Description</td>
<td>List of error, warning, info or other validation messages returned by SGML parser</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### ErrorLevelMessageCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of error-level validation messages returned by SGML parser</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### WarningLevelMessageCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of warning-level validation messages returned by SGML parser</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>

### InfoLevelMessageCount Property

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>int</td>
</tr>
<tr>
<td>Description</td>
<td>Count of info-level validation messages returned by SGML parser</td>
</tr>
<tr>
<td>Reference</td>
<td></td>
</tr>
</tbody>
</table>
8 Configuration

Configuring the OpenSp Parser

Please see the User Guide for information on specifying executable path (and any catalog file path) for the OpenSP executable. If you are using the UNIX/Cygwin version of OpenSP, then you can specify the catalog path in bean “SgmlCatalogPath”. If you are using the Windows version of OpenSp, then you will be editing the bean “WindowsSgmlCatalogPath”.

---

<table>
<thead>
<tr>
<th>Property</th>
<th>Identifier</th>
<th>Type</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>QnameLevelMessageCount Property</td>
<td><a href="http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument">http://jhove2.org/terms/property/org/jhove2/module/format/sgml/SgmlDocument</a> Properties/QnameLevelMessageCount</td>
<td>int</td>
<td>Count of quantity (name-length) validation messages returned by SGML parser</td>
<td></td>
</tr>
</tbody>
</table>
Configuring ONSGMLS options

Options for the OpenSP onsgmls executable (which parses and validates the SGML file) are specified in the Spring Configuration file ($JHOVE_HOME)/config/spring/module/format/sgml/jhove2-sgml-config.xml. If you are using the UNIX/Cygwin version of OpenSP, then you will be editing bean “OnsgmlsOptions”. If you are using the Windows version of OpenSp, then you will be editing the bean “WindowsOnsgmlsOptions”.

The “filepathFilter” property of these beans refer to the bean which specifies the appropriate Java class to filter the input file name, and all temporarily created files, into a platform-specific format.

Please see Appendix A: OpenSp ONSGMLS Options for details about each of the configuration options for the onsgmls executable.

Configuring OSGMLNorm Options

Options for the OpenSP osgmlnorm executable (which parses the SGML file and extracts the DOCTYPE statement) are specified in the Spring Configuration file ($JHOVE_HOME)/config/spring/module/format/sgml/jhove2-sgml-config.xml. If you are using the UNIX/Cygwin version of OpenSP, then you will be editing bean “SgmlNormOptions”. If you are using the Windows version of OpenSp, then you will be editing the bean “WindowsSgmlNormOptions”.

The “filepathFilter” property of these beans refer to the bean which specifies the appropriate Java class to filter the input file name, and all temporarily created files, into a platform-specific format.

If you do not need the features of the DOCTYPE statement (the public and system identifiers), then you can turn off execution of this tool by setting the “shouldFindDoctype” property in the SGMLModule bean to “false”.

Please see Appendix B OpenSP OSGMLNORM Options for details about each of the configuration options for the OpenSp osgmlnorm executable.

9 Implementation Notes

Wrapping OpenSP

The use of OpenSP to parse and validate SGML is wrapped by the class org.jhove2.module.format.sgml.OpenSpWrapper, which implements the interface org.jhove2.module.format.sgml.SgmlParser. It is configured with an instance of an org.jhove2.util.externalprocess.ExternalProcessHandler, to manage platform-specific invocation of process external to the Java JVM (as onsgmls and osgmlnorm both are). It is also configured with a org.jhove2.util.externalprocess.FilePathFilter object to handle platform-specific formatting of file paths (file path separators, etc.).
After invoking the OpenSp utilities, OpenSpWrapper invokes the wrappers around the ANLTR-generated lexers and parsers for the output and error files of these utilities, to extract significant properties (including error, warning, and info messages) about the file.

**External Process Handler**

The package `org.jhove2.util.externalprocess` contains utility classes for invoking external processes in a platform-appropriate manner via an invocation of the Java Runtime.exec() method.

Users should be aware that implementation of this fork is platform specific, and can result in the creation of second process whose memory footprint, at least initially, is equal in size to that of the forking application.

**The SGML Catalog**

Unless the SGML files you are processing are known to contain system identifier that directly and correctly resolve to a location on your system, or to a reachable URL, you will find it necessary to create an SGML catalog file to resolve public or system identifiers to the actual location of any DTD required by any SGML files you are characterizing. Please see the sections above on configuring the path to the catalog in the Spring configuration file for this module.

**Using ANTLR to Parse ESIS output**

JHOVE2 contains three ANTLR grammar files for parsing 3 types of output from OpenSP:

- **ESISCommands.g** is the grammar for parsing ESIS output from ongmls.
- **DoctypeFinder.g** is the grammar for parsing output from osgmlnorm.
- **SgmlParseMessages.g** is the grammar for parsing error output from both utilities

These grammars can be found in the source distribution in directory `src/main/antlr3/org/jhove2/module/format/sgml`. The Maven pom.xml file specifies the ANTLR Maven plugin to automatically generate Java classes from these grammar files at compile-time. Maven places these generated Java classes in directory `target/generated-sources/antlr3/org/jhove2/module/format/sgml`.

These classes are in turn wrapped by 3 corresponding classes, which invoke the parsers, and extract the significant property information from the parse:

- **EsisParser**
- **DoctypeParser**
- **OpenSpMessageParser**
These classes can be found in package org.jhove2.module.format.sgml.

Appendix A: OpenSP ONSGMLS Options

The following is reproduced from the documentation provided in the source distribution of OpenSP (Copyright (c) 1994, 1995, 1996, 1997, 1998 James Clark).

The home page for OpenSP and OpenJade can be found at http://openjade.sourceforge.net

ONSGMLS

An SGML System Conforming to International Standard ISO 8879 -- Standard Generalized Markup Language

An SGML Extended Facilities system conforming to Annex A of International Standard ISO/IEC 10744 -- Hypermedia/Time-based Structuring Language

SYNOPSIS

onsgmls [option...] sysid...

DESCRIPTION

Onsgmls parses and validates the SGML document whose document entity is specified by the system identifiers sysid... and prints on the standard output a simple text representation of its Element Structure Information Set. (This is the information set which a structure-controlled conforming SGML application should act upon.) If more than one system identifier is specified, then the corresponding entities will be concatenated to form the document entity. Thus the document entity may be spread amongst several files; for example, the SGML declaration, prolog and document instance set could each be in a separate file. If no system identifiers are specified, then nsgmls will read the document entity from the standard input. A command line system identifier of – can be used to refer to the standard input. (Normally in a system identifier, <osfd>0 is used to refer to standard input.)

OPTIONS

The following options are available:

-aname, --activate=name

Make link type or doctype name active. Not all ESIS information is output in this case: the active LPDs are not explicitly reported, although each link attribute is qualified with its link type name; there is no information about result elements; when there are multiple link rules applicable to
the current element, nsgmls always chooses the first. You cannot mix active link types and doctypes. Only one active doctype can be specified.

- **bname**, **--bctf=name**
- **bname**, **--encoding=name**

This determines the encoding used for output. If in fixed character set mode it specifies the name of an encoding; if not, it specifies the name of a BCTF.

- **B**, **--batch-mode**

Batch mode. Parse each **sysid...** specified on the command line separately, rather than concatenating them. This is useful mainly with -s.

If **-tfilename** is also specified, then the specified **filename** will be prefixed to the **sysid** to make the filename for the RAST result for each **sysid**.

- **csysid**, **--catalog=sysid**

Map public identifiers and entity names to system identifiers using the catalog entry file whose system identifier is **sysid**. Multiple -c options are allowed. If there is a catalog entry file called catalog in the same place as the document entity, it will be searched for immediately after those specified by -c.

- **C**, **--catalogs**

The **filename** arguments specify catalog files rather than the document entity. The document entity is specified by the first DOCUMENT entry in the catalog files.

- **D**directory, **--directory=directory**

Search directory for files specified in system identifiers. Multiple -D options are allowed. See the description of the osfile storage manager for more information about file searching.

- **e**, **--open-entities**

Describe open entities in error messages. Error messages always include the position of the most recently opened external entity.

- **E**number, **--max-errors=number**

Onsgmls will exit after **number** errors. If **number** is 0, there is no limit on the number of errors. The default is 200.

- **f**file, **--error-file=file**
Redirect errors to file. This is useful mainly with shells that do not support redirection of stderr.

-g, --open-elements

Show the generic identifiers of open elements in error messages.

-h, --help

Show a help message and exit.

-iname, --include=name

Pretend that

<!ENTITY % name "INCLUDE">

occurs at the start of the document type declaration subset in the SGML document entity. Since repeated definitions of an entity are ignored, this definition will take precedence over any other definitions of this entity in the document type declaration. Multiple -i options are allowed. If the SGML declaration replaces the reserved name INCLUDE then the new reserved name will be the replacement text of the entity. Typically the document type declaration will contain

<!ENTITY % name "IGNORE">

and will use %name; in the status keyword specification of a marked section declaration. In this case the effect of the option will be to cause the marked section not to be ignored.

-n, --error-numbers

Show message numbers in error messages.

-ooption, --option=option

Output additional information according to option:

entity

Output definitions of all general entities not just for data or subdoc entities that are referenced or named in an ENTITY or ENTITIES attribute.

id

Distinguish attributes whose declared value is ID.

line

Output L commands giving the current line number and filename.

included

Output an i command for included subelements.
empty

Output an e command for elements which are not allowed to have an end-tag, that is those with a declared content of empty or with a content reference attribute.

notation-sysid

Output an f command before an N command, if a system identifier could be generated for that notation.

nonsigml

In fixed character set mode, output \% escape sequences for non-SGML data characters. Non-SGML data characters can result from numeric character references.

data-attribute

Output the notation name and attributes for DATA attributes. Otherwise, DATA attributes are treated like CDATA attributes. For more details, see clause 4.4.3 of Annex K of ISO 8879.

comment

Output an _ command with the contents of a comment. Multiple comments in a single comment declaration will result in multiple distinct _ commands, just as if the comments were each in a separate comment declaration.

omitted

Output an o command before a command which was implied by the input document, but omitted from the actual markup. This currently affects (, ), and A commands.

tagomit

As omitted, but only for ( and ) commands.

attromit

As omitted, but only for A commands.

Multiple -o options are allowed.

-p, --only-prolog

Parse only the prolog. Onsgmls will exit after parsing the document type declaration. Implies -s.

-R, --restricted

Restrict file reading. This option is intended for use with onsgmls-based Web tools (e.g., CGI scripts) to prevent reading of arbitrary files on the Web server. With this option enabled, onsgmls will not read any local files unless they are located in a directory (or subdirectory)
specified by the -D option or included in the SGML_SEARCH_PATH environment variable. As a further security precaution, this option limits filenames to the characters A-Z, a-z, 0-9, '/', '.', '_', \ and does not allow filenames containing ".".

-s, --no-output

Suppress output. Error messages will still be printed.

-tfile, --rast-file=file

Output to file the RAST result as defined by ISO/IEC 13673:1995 (actually this isn't quite an IS yet; this implements the Intermediate Editor's Draft of 1994/08/29, with changes to implement ISO/IEC JTC1/SC18/WG8 N1777). The normal output is not produced.

-v, --version

Print the version number.

-wtype, --warning=type

Control warnings and errors. Multiple -w options are allowed. The following values of type enable warnings:

xml

Warn about constructs that are not allowed by XML. The warnings can be enabled individually.
mixed

Warn about mixed content models that do not allow #PCDATA anywhere.

sgmldecl

Warn about various dubious constructions in the SGML declaration.

should

Warn about various recommendations made in ISO 8879 that the document does not comply with. (Recommendations are expressed with "should", as distinct from requirements which are usually expressed with "shall".)

default

Warn about defaulted references.
duplicate

Warn about duplicate entity declarations.

undefined
Warn about undefined elements: elements used in the DTD but not defined.

unclosed

Warn about unclosed start and end-tags.

empty

Warn about empty start and end-tags.

net

Warn about net-enabling start-tags and null end-tags.

min-tag

Warn about minimized start and end-tags. Equivalent to combination of unclosed, empty and net warnings.

unused-map

Warn about unused short reference maps: maps that are declared with a short reference mapping declaration but never used in a short reference use declaration in the DTD.

unused-param

Warn about parameter entities that are defined but not used in a DTD. Unused internal parameter entities whose text is INCLUDE or IGNORE won't get the warning.

notation-sysid

Warn about notations for which no system identifier could be generated.

all

Warn about conditions that should usually be avoided (in the opinion of the author). Equivalent to: mixed, should, default, undefined, sgmldecl, unused-map, unused-param, empty and unclosed.

immediate-recursion

Warn about immediately recursive elements. For more details, see clause 2.2.5 of Annex K of ISO 8879.

fully-declared

Warn if the document instance fails to be fully-declared. This has the effect of changing the SGML declaration to specify IMPLYDEF ATTLIST NO ELEMENT NO ENTITY NO NOTATION NO. For more details, see clause 2.2.1 of Annex K of ISO 8879.

fully-tagged
Warn if the document instance fails to be fully-tagged. This has the effect of changing the SGML declaration to specify DATATAG NO, RANK NO, OMITTAG NO, SHORTTAG STARTTAG EMPTY NO and SHORTTAG ATTRIB OMITNAME NO. For more details, see clause 2.2.2 of Annex K of ISO 8879.

Warn if the document instance fails to be amply-tagged. Implicitly defined elements may be immediately recursive if -recursive is specified. This has the effect of changing the SGML declaration to specify DATATAG NO, RANK NO, OMITTAG NO, SHORTTAG ATTRIB OMITNAME NO and either IMPLYDEF ELEMENT ANYOTHER or IMPLYDEF ELEMENT YES. For more details, see clause 2.2.4 of Annex K of ISO 8879.

type-valid

Warn if the document instance fails to be type-valid. This has the effect of changing the SGML declaration to specify VALIDITY YES. For more details, see clause 2.2.3 of Annex K of ISO 8879.

entity-ref

Warn about references to non-predefined entities. This has the effect of changing the SGML declaration to specify ENTITIES REF NONE. For more details, see clause 2.3.2 of Annex K of ISO 8879.

external-entity-ref

 Warn about references to external entities. This includes references to an external DTD subset. This has the effect of changing the SGML declaration to specify ENTITIES REF INTERNAL. For more details, see clause 2.3.3 of Annex K of ISO 8879.

integral

Warn if the document instance is not integrally stored. This has the effect of changing the SGML declaration to specify ENTITIES INTEGRAL YES. For more details, see clause 2.3.1 of Annex K of ISO 8879.

A warning can be disabled by using its name prefixed with no-. Thus no-wall -wno-duplicate will enable all warnings except those about duplicate entity declarations.

The following values for warning_type disable errors:

no-idref

Do not give an error for an ID reference value which no element has as its ID. The effect will be as if each attribute declared as an ID reference value had been declared as a name.
no-significant

Do not give an error when a character that is not a significant character in the reference concrete syntax occurs in a literal in the SGML declaration. This may be useful in conjunction with certain buggy test suites.

no-valid

Do not require the document to be type-valid. This has the effect of changing the SGML declaration to specify `VALIDITY NOASSERT` and `IMPLYDEF ATTLIST YES ELEMENT YES ENTITY YES NOTATION YES`. An option of `-wvalid` has the effect of changing the SGML declaration to specify `VALIDITY TYPE` and `IMPLYDEF ATTLIST NO ELEMENT NO ENTITY NO NOTATION NO`. If neither `-wvalid` nor `-wno-valid` are specified, then the `VALIDITY` and `IMPLYDEF` specified in the SGML declaration will be used.

no-afdr

Do not give errors when AFDR meta-DTD notation features are used in the DTD. These errors are normally produced when parsing the DTD, but suppressed when parsing meta-DTDs.

-x, --references

Show information about relevant clauses (from ISO 8879:1986) in error messages.

The following options are also supported for backwards compatibility with sgmls:

-d

Same as `-wduplicate`.

-l

Same as `-oline`.

-msysid

Same as `-c`.

-r

Same as `-wdefault`.

-u

Same as `-wundef`.
Appendix B: OpenSP OSGMLNORM Options

The following is reproduced from the documentation provided in the source distribution of OpenSP (Copyright (c) 1994, 1995, 1996, 1997, 1998 James Clark).

The home page for OpenSP and OpenJade can be found at http://openjade.sourceforge.net

OSGMLNORM

An SGML System Conforming to International Standard ISO 8879 -- Standard Generalized Markup Language

An SGML Extended Facilities system conforming to Annex A of International Standard ISO/IEC 10744 -- Hypermedia/Time-based Structuring Language

SYNOPSIS

osgmlnorm [option...] sysid...

DESCRIPTION

Osgmlnorm prints on the standard output a normalized document instance for the SGML document contained in the concatenation of the entities with system identifiers sysid....

When the normalized instance is prefixed with the original SGML declaration and prolog, it will have the same ESIS as the original SGML document, with the following exceptions:

- The output of osgmlnorm does not protect against the recognition of short reference delimiters, so any USEMAP declarations must be removed from the DTD.
- The normalized instance will use the reference delimiters, even if the original instance did not.
- If marked sections are included in the output using the -m option, the reference reserved names will be used for the status keywords even if the original instance did not.
- Any ESIS information relating to the SGML LINK feature will be lost.

The normalized instance will not use any markup minimization features except that:

- Any attributes that were not specified in the original instance will not be included in the normalized instance. (Current attributes will be included.)
- If the declared value of an attribute was a name token group, and a value was specified that was the same as the name of the attribute, then the attribute name and value indicator will be omitted. For example, with HTML osgmlnorm would output <DL COMPACT> rather than <DL COMPACT="COMPACT">
**OPTIONS**

The following options are available:

- `bname, --bctf=name`
  
  Use the **BCTF** *name* for output.

- `csysid, --catalog=sysid`
  
  Map public identifiers and entity names to system identifiers using the catalog entry file whose system identifier is *sysid*. This has the same effect as in `onsgmls`.

- `C, --catalogs`
  
  This has the same effect as in `onsgmls`.

- `d, --dtd`
  
  Output a document type declaration with the same external identifier as the input document, and with no internal declaration subset. No check is performed that the document instance is valid with respect to this DTD.

- `D directory, --directory=directory`
  
  Search *directory* for files specified in system identifiers. This has the same effect as in `onsgmls`.

- `e, --open-entities`
  
  Describe open entities in error messages.

- `iname, --include=name`
  
  This has the same effect as in `onsgmls`.

- `m, --marked-sections`
  
  Output any marked sections that were in the input document instance.

- `n, --comments`
  
  Output any comments that were in the input document instance.

- `r, --raw`
  
  Raw output. Don't perform any conversion on RSs and REs when printing the entity. The entity would typically have the storage manager attribute `records=asis`. 

JHOVE2: Next-Generation Architecture for Format-Aware Characterization
-R, --restricted

   This has the same effect as in onsgmls.

-v, --version

   Print the version number.

-wtype, --warning=type

   Control warnings and errors according to type. This has the same effect as in onsgmls.

Appendix C: OpenSP Output Format (ESIS)

The following is reproduced from the documentation provided in the source distribution of OpenSP (Copyright (c) 1994, 1995, 1996, 1997, 1998 James Clark).

The home page for OpenSP and OpenJade can be found at http://openjade.sourceforge.net

Onsgmls Output Format

The output is a series of lines. Lines can be arbitrarily long. Each line consists of an initial command character and one or more arguments. Arguments are separated by a single space, but when a command takes a fixed number of arguments the last argument can contain spaces. There is no space between the command character and the first argument. Arguments can contain the following escape sequences:

\ 
   A \.

\n
   A record end character.

\|

   Internal SDATA entities are bracketed by these.

\nnn

   The character whose code is nnn octal.
A record start character will be represented by \012. Most applications will need to ignore \012 and translate \n into newline.

\#n;

The character whose number is \n decimal in the internal character set. \n can have any number of digits. This is used for characters that are not representable by the encoding used for output. This will only occur with the multibyte version of nsgmls.

\%n;

The character whose number is \n decimal in the document character set. \n can have any number of digits. This is used for data characters resulting from numeric character references to non-SGML characters in fixed character set mode.

The possible command characters and arguments are as follows:

\(gi\)

The start of an element whose generic identifier is gi. Any attributes for this element will have been specified with A commands.

\)gi

The end of an element whose generic identifier is gi.

-data

Data.

&name

A reference to an external data entity name; name will have been defined using an E command.

*pi

A processing instruction with data pi.

Aname val

The next element to start has an attribute name with value val which takes one of the following forms:

IMPLIED

The value of the attribute is implied.

CDATA data

The attribute is character data. This is used for attributes whose declared value is CDATA.
NOTATION nname

The attribute is a notation name; nname will have been defined using a N command. This is used for attributes whose declared value is NOTATION.

ENTITY name...

The attribute is a list of general entity names. Each entity name will have been defined using an I, E or S command. This is used for attributes whose declared value is ENTITY or ENTITIES.

TOKEN token...

The attribute is a list of tokens. This is used for attributes whose declared value is anything else.

ID token

The attribute is an ID value. This will be output only if the -oid option is specified. Otherwise TOKEN will be used for ID values.

DATA nname data

The attribute is character data with an associated notation. The definition of the notation and any applicable attributes of the notation will be given using subsequent Dname val lines. This is used for attributes whose declared value is DATA. It will be output only if the -odata-attribute option is specified. Otherwise CDATA will be used for DATA values.

Dename name val

This is the same as the A command, except that it specifies a data attribute for an external entity named ename. Any D commands will come after the E command that defines the entity to which they apply, but before any & or A commands that reference the entity.

atypemanameval

The next element to start has a link attribute with link type type, name name, and value val, which takes the same form as with the A command.

Nnname

Define a notation nname. This command will be preceded by a p command if the notation was declared with a public identifier, and by a s command if the notation was declared with a system identifier. If the -onotation-sysid option was specified, this command will also be preceded by an f command giving the system identifier generated by the entity manager (unless it was unable to generate one). A notation will only be defined if it is to be referenced in an E command or in an A command for an attribute with a declared value of NOTATION.

Eename typ nname
Define an external data entity named `ename` with type `typ` (CDATA, NDATA or SDATA) and notation `not`. This command will be preceded by an `f` command giving the system identifier generated by the entity manager (unless it was unable to generate one), by a `p` command if a public identifier was declared for the entity, and by a `s` command if a system identifier was declared for the entity. `not` will have been defined using a `N` command. Data attributes may be specified for the entity using `D` commands. If the `-oentity` option is not specified, an external data entity will only be defined if it is to be referenced in a `&` command or in an `A` command for an attribute whose declared value is `ENTITY` or `ENTITIES`.

```lisp
I name typ text
```

Define an internal data entity named `ename` with type `typ` and entity text `text`. The `typ` will be CDATA or SDATA unless the `-oentity` option was specified, in which case it can also be PI or TEXT (for an SGML text entity). If the `-oentity` option is not specified, an internal data entity will only be defined if it is referenced in an `A` command for an attribute whose declared value is `ENTITY` or `ENTITIES`.

```lisp
S name
```

Define a subdocument entity named `ename`. This command will be preceded by an `f` command giving the system identifier generated by the entity manager (unless it was unable to generate one), by a `p` command if a public identifier was declared for the entity, and by a `s` command if a system identifier was declared for the entity. If the `-oentity` option is not specified, a subdocument entity will only be defined if it is referenced in a `{` command or in an `A` command for an attribute whose declared value is `ENTITY` or `ENTITIES`.

```lisp
T name
```

Define an external SGML text entity named `ename`. This command will be preceded by an `f` command giving the system identifier generated by the entity manager (unless it was unable to generate one), by a `p` command if a public identifier was declared for the entity, and by a `s` command if a system identifier was declared for the entity. This command will be output only if the `-oentity` option is specified.

```lisp
ssysid
```

This command applies to the next `E`, `S`, `T` or `N` command and specifies the associated system identifier.

```lisp
ppubid
```

This command applies to the next `E`, `S`, `T` or `N` command and specifies the associated public identifier.

```lisp
fsysid
```
This command applies to the next E, S, T or, if the -notation-sysid option was specified, N command and specifies the system identifier generated by the entity manager from the specified external identifier and other information about the entity or notation.

{ename

The start of the SGML subdocument entity ename; ename will have been defined using a S command.

}ename

The end of the SGML subdocument entity ename.

Llineno file

Llineno

Set the current line number and filename. The file argument will be omitted if only the line number has changed. This will be output only if the -l option has been given.

#text

An APPINFO parameter of text was specified in the SGML declaration. This is not strictly part of the ESIS, but a structure-controlled application is permitted to act on it. No # command will be output if APPINFO NONE was specified. A # command will occur at most once, and may be preceded only by a single L command.

C

This command indicates that the document was a conforming SGML document. If this command is output, it will be the last command. An SGML document is not conforming if it references a subdocument entity that is not conforming.

i

The next element to start is an included subelement. This will be output only if the -oincluded option is specified.

e

The next element to start has a declared content of EMPTY or a content reference attribute, and so its end-tag must be omitted. This will be output only if the -oempty option is specified.

_comment

A comment with data comment. This will be output only if the -ocomment option is specified.
The actual markup for the next \( \), or \( A \) command was omitted from the input. This will be output only if one of the \(-\text{oomitted}, -\text{otagomit}, \text{or } -\text{oattromit} \) options is specified.