SiSU
Commands

Ralph Amissah

* Generated by SiSU www.jus.uio.no/sisu
www.sisudoc.org

Document information:
sourcefile sisu_introduction.sst
Generated by SiSU www.jus.uio.no/sisu
version information: SiSU 0.59.1 of 2007w39/2

For alternative output formats of this document check:
<http://www.jus.uio.no/sisu/sisu_introduction/sisu_manifest.html>

Contents

? DESCRIPTION

1. Introduction - What is SiSU?

SiSU is a system for document markup, publishing (in multiple open standard formats) and search

SiSU [1] is a[2] framework for document structuring, publishing and search, comprising of (a) a lightweight document structure and presentation markup syntax and (b) an accompanying engine for generating standard document format outputs from documents prepared in sisu markup syntax, which is able to produce multiple standard outputs that (can) share a common numbering system for the citation of text within a document.

SiSU is developed under an open source, software libre license (GPL3). It has been developed in the context of coping with large document sets with evolving markup related technologies, for which you want multiple output formats, a common mechanism for cross-output-format citation, and search.

SiSU both defines a markup syntax and provides an engine that produces open standards format outputs from documents prepared with **SiSU** markup. From a single lightly prepared document sisu custom builds several standard output formats which share a common (text object) numbering system for citation of content within a document (that also has implications for search). The sisu engine works with an abstraction of the document's structure and content from which it is possible to generate different forms of representation of the document. Significantly **SiSU** markup is more sparse than html and outputs which include html, LaTeX, landscape and portrait pdfs, Open Document Format (ODF), all of which can be added to and updated. **SiSU** is also able to populate SQL type databases at an object level, which means that searches can be made with that degree of granularity. Results of objects (primarily paragraphs and

[1] "**SiSU** information Structuring Universe" or "Structured information, Serialized Units". also chosen for the meaning of the Finnish term "sisu".

[2] Unix command line oriented

headings) can be viewed directly in the database, or just the object numbers shown - your search criteria is met in these documents and at these locations within each document.

9 Source document preparation and output generation is a two step process: (i) document source is prepared, that is, marked up in sisu markup syntax and (ii) the desired output subsequently generated by running the sisu engine against document source. Output representations if updated (in the sisu engine) can be generated by re-running the engine against the prepared source. Using **SiSU** markup applied to a document, **SiSU** custom builds various standard open output formats including plain text, HTML, XHTML, XML, OpenDocument, LaTeX or PDF files, and populate an SQL database with objects[3] (equating generally to paragraph-sized chunks) so searches may be performed and matches returned with that degree of granularity (e.g. your search criteria is met by these documents and at these locations within each document). Document output formats share a common object numbering system for locating content. This is particularly suitable for "published" works (finalized texts as opposed to works that are frequently changed or updated) for which it provides a fixed means of reference of content.

10 In preparing a **SiSU** document you optionally provide semantic information related to the document in a document header, and in marking up the substantive text provide information on the structure of the document, primarily indicating heading levels and footnotes. You also provide information on basic text attributes where used. The rest is automatic, sisu from this information custom builds[4] the different forms of output requested.

11 **SiSU** works with an abstraction of the document based on its structure which is comprised of its frame[5] and the objects[6] it contains, which enables **SiSU** to represent the document in many different ways, and to take advantage of the strengths of different ways of presenting documents. The objects are numbered, and these numbers can be used to provide a common base for citing material within a document across the different output format types. This is significant as page numbers are not suited to the digital age, in web publishing, changing a browser's default font or using a different browser means that text appears on different pages; and in publishing in different formats, html, landscape and portrait pdf etc. again page numbers are of no use to cite text in a manner that is relevant against the different output types. Dealing with documents at an object level together with object numbering also has implications for search.

12 One of the challenges of maintaining documents is to keep them in a format that would allow users to use them without depending on a proprietary software popular at the time. Consider the ease of dealing with legacy proprietary formats today and what guarantee you have that old proprietary formats will remain (or can be read without proprietary software/equipment) in 15 years time, or the way the way in which html has evolved over its relatively short span of existence. **SiSU** provides the flexibility of outputing documents in multiple non-proprietary open formats including html, pdf[7] and the ISO standard ODF.[8] Whilst **SiSU** relies on software, the markup is uncomplicated and minimalistic which guarantees that future engines can be written to run against it. It is also easily converted to other formats, which means documents prepared in **SiSU** can be migrated to other document formats. Further security is provided by the fact that the software itself, **SiSU** is available under GPL3 a licence that guarantees that the source code will always be open, and free as in libre which means that that code base can be used updated and further

[3] objects include: headings, paragraphs, verse, tables, images, but not footnotes/endnotes which are numbered separately and tied to the object from which they are referenced.

[4] i.e. the html, pdf, odf outputs are each built individually and optimised for that form of presentation, rather than for example the html being a saved version of the odf, or the pdf being a saved version of the html.

[5] the different heading levels

[6] units of text, primarily paragraphs and headings, also any tables, poems, code-blocks

[7] Specification submitted by Adobe to ISO to become a full open ISO specification <http://www.linux-watch.com/news/NS7542722606.html>

[8] ISO/IEC 26300:2006

developed as required under the terms of its license. Another challenge is to keep up with a moving target. **SiSU** permits new forms of output to be added as they become important, (Open Document Format text was added in 2006), and existing output to be updated (html has evolved and the related module has been updated repeatedly over the years, presumably when the World Wide Web Consortium (w3c) finalises html 5 which is currently under development, the html module will again be updated allowing all existing documents to be regenerated as html 5).

13 The document formats are written to the file-system and available for indexing by independent indexing tools, whether off the web like Google and Yahoo or on the site like Lucene and Hyperestraier.

14 **SiSU** also provides other features such as concordance files and document content certificates, and the working against an abstraction of document structure has further possibilities for the research and development of other document representations, the availability of objects is useful for example for topic maps and the commercial law thesaurus by Vikki Rogers and Al Krtizer, together with the flexibility of **SiSU** offers great possibilities.

15 **SiSU** is primarily for published works, which can take advantage of the citation system to reliably reference its documents. **SiSU** works well in a complementary manner with such collaborative technologies as Wikis, which can take advantage of and be used to discuss the substance of content prepared in **SiSU** .

16 <http://www.jus.uio.no/sisu>

2. How does sisu work?

17

18 **SiSU** markup is fairly minimalistic, it consists of: a (largely optional) document header, made up of information about the document (such as when it was published, who authored it, and granting what rights) and any processing instructions; and markup within the substantive text of the document, which is related to document structure and typeface. **SiSU** must be able to discern the structure of a document, (text headings and their levels in relation to each other), either from information provided in the document header or from markup within the text (or from a combination of both). Processing is done against an abstraction of the document comprising of information on the document's structure and its objects,[2] which the program serializes (providing the object numbers) and which are assigned hash sum values based on their content. This abstraction of information about document structure, objects, (and hash sums), provides considerable flexibility in representing documents different ways and for different purposes (e.g. search, document layout, publishing, content certification, concordance etc.), and makes it possible to take advantage of some of the strengths of established ways of representing documents, (or indeed to create new ones).

3. Summary of features

- sparse/minimal markup (clean utf-8 source texts). Documents are prepared in a single UTF-8 file using a minimalistic mnemonic syntax. Typical literature, documents like “War and Peace” require almost no markup, and most of the headers are optional.

- markup is easily readable/parsable by the human eye, (basic markup is simpler and more sparse than the most basic HTML), [this may also be converted to XML representations of the same input/source document].

- markup defines document structure (this may be done once in a header pattern-match description, or for heading levels individually); basic text attributes (bold, italics, underscore, strike-through etc.) as required; and semantic information related to the document (header information, extended beyond the Dublin core and easily further extended as required); the headers may also contain processing instructions. **SiSU** markup is primarily an abstraction of document structure and document metadata to permit taking advantage of the basic strengths of existing alternative practical standard ways of representing documents [be that browser viewing, paper publication, sql search etc.] (html, xml, odf, latex, pdf, sql)

- for output produces reasonably elegant output of established industry and institutionally accepted open standard formats.[3] takes advantage of the different strengths of various standard formats for representing documents, amongst the output formats currently supported are:

    - html - both as a single scrollable text and a segmented document
    - xhtml
    - XML - both in sax and dom style xml structures for further development as required
    - ODF - open document format, the iso standard for document storage
    - LaTeX - used to generate pdf
    - pdf (via LaTeX)
    - sql - population of an sql database, (at the same object level that is used to cite text within a document)

Also produces: concordance files; document content certificates (md5 or sha256 digests of headings, paragraphs, images etc.) and html manifests (and sitemaps of content). (b) takes advantage of the strengths implicit in these very different output types, (e.g. PDFs produced using typesetting of LaTeX, databases populated with documents at an individual object/paragraph level, making possible granular search (and related possibilities))

- ensuring content can be cited in a meaningful way regardless of selected output format. Online publishing (and publishing in multiple document formats) lacks a useful way of citing text internally within documents (important to academics generally and to lawyers) as page numbers are meaningless across browsers and formats. sisu seeks to provide a common way of pinpoint the text within a document, (which can be utilized for citation and by search engines). The outputs share a common numbering system that is meaningful (to man and machine) across all digital outputs whether paper, screen, or database oriented, (pdf, HTML, xml, sqlite, postgresql), this numbering system can be used to reference content.

- Granular search within documents. SQL databases are populated at an object level (roughly headings, paragraphs, verse, tables) and become searchable with that degree of granularity, the output information provides the object/paragraph numbers which are relevant across all generated outputs; it is also possible to look at just the matching paragraphs of the documents in the database; [output indexing also work well with search indexing tools like hyperestraier].

- long term maintainability of document collections in a world of changing formats, having a very sparsely marked-up source document base.

there is a considerable degree of future-proofing, output representations are "upgradeable", and new document formats may be added. e.g. addition of odf (open document text) module in 2006 and in future html5 output sometime in future, without modification of existing prepared texts

- SQL search aside, documents are generated as required and static once generated.
- documents produced are static files, and may be batch processed, this needs to be done only once but may be repeated for various reasons as desired (updated content, addition of new output formats, updated technology document presentations/representations)
- document source (plaintext utf-8) if shared on the net may be used as input and processed locally to produce the different document outputs
- document source may be bundled together (automatically) with associated documents (multiple language versions or master document with inclusions) and images and sent as a zip file called a sisupod, if shared on the net these too may be processed locally to produce the desired document outputs
- generated document outputs may automatically be posted to remote sites.
- for basic document generation, the only software dependency is **Ruby** , and a few standard Unix tools (this covers plaintext, HTML, XML, ODF, LaTeX). To use a database you of course need that, and to convert the LaTeX generated to pdf, a latex processor like tetex or texlive.
- as a developers tool it is flexible and extensible

Syntax highlighting for **SiSU** markup is available for a number of text editors.

SiSU is less about document layout than about finding a way with little markup to be able to construct an abstract representation of a document that makes it possible to produce multiple representations of it which may be rather different from each other and used for different purposes, whether layout and publishing, or search of content

i.e. to be able to take advantage from this minimal preparation starting point of some of the strengths of rather different established ways of representing documents for different purposes, whether for search (relational database, or indexed flat files generated for that purpose whether of complete documents, or say of files made up of objects), online viewing (e.g. html, xml, pdf), or paper publication (e.g. pdf)...

the solution arrived at is by extracting structural information about the document (about headings within the document) and by tracking objects (which are serialized and also given hash values) in the manner described. It makes possible representations that are quite different from those offered at present. For example objects could be saved individually and identified by their hashes, with an index of how the objects relate to each other to form a document.

DOCUMENT INFORMATION (METADATA)

Metadata

Document Manifest @

<http://www.jus.uio.no/sisu/sisu_manual/sisu_introduction/sisu_manifest.html>

Dublin Core (DC)

DC tags included with this document are provided here.

DC Title: SiSU - Commands

DC Creator: Ralph Amissah

DC Rights: Copyright (C) Ralph Amissah 2007, part of SiSU documentation, License GPL 3

DC Type: information

DC Date created: 2002-08-28

DC Date issued: 2002-08-28

DC Date available: 2002-08-28

DC Date modified: 2007-09-16

DC Date: 2007-09-16

Version Information

Sourcefile: sisu_introduction.sst

Filetype: SiSU text 0.58

Sourcefile Digest, MD5(sisu_introduction.sst)= 877333106803c1fc864bccdbd0c667e2

Skin_Digest: MD5(/home/ralph/grotto/theatre/dbld/builds/sisu/sisu/data/doc/sisu/sisu_markup_samples/sisu_manual/_sisu/skin/doc/skin_sisu_manual.rb)= 20fc43cf3eb6590bc3399a1aef65c5a9

Generated

Document (metaverse) last generated: Tue Sep 25 02:52:52 +0100 2007

Generated by: SiSU 0.59.1 of 2007w39/2 (2007-09-25)

Ruby version: ruby 1.8.6 (2007-06-07 patchlevel 36) [i486-linux]

Information on this document copy and an unofficial List of Some web related information and sources

”Support Open Standards and Software Libre for the Information Technology Infrastructure” RA

Information on this document copy www.jus.uio.no/sisu/

W3 since October 3 1993 SiSU 1997, current 2007.
SiSU presentations at **www.jus.uio.no/sisu/**

SiSU **pdf** versions can be found at:
http://www.jus.uio.no/sisu/sisu_introduction/portrait.pdf
http://www.jus.uio.no/sisu/sisu_introduction/landscape.pdf

SiSU **html** versions may be found at:
http://www.jus.uio.no/sisu/sisu_introduction/toc.html or
http://www.jus.uio.no/sisu/sisu_introduction/doc.html

SiSU Manifest of document output and metadata may be found at:
http://www.jus.uio.no/sisu/sisu_introduction/sisu_manifest.html

SiSU found at: **www.jus.uio.no/sisu/**

Links that may be of interest at SiSU and elsewhere:

SiSU Manual
http://www.jus.uio.no/sisu/sisu_manual/

Book Samples and Markup Examples
http://www.jus.uio.no/sisu/SiSU/2.html

SiSU @ Wikipedia
http://en.wikipedia.org/wiki/SiSU

SiSU @ Freshmeat
http://freshmeat.net/projects/sisu/

SiSU @ Ruby Application Archive
http://raa.ruby-lang.org/project/sisu/

SiSU @ Debian
http://packages.qa.debian.org/s/sisu.html

SiSU Download
http://www.jus.uio.no/sisu/SiSU/download.html

SiSU Changelog
http://www.jus.uio.no/sisu/SiSU/changelog.html

SiSU help
http://www.jus.uio.no/sisu/sisu_manual/sisu_help/

SiSU help sources
http://www.jus.uio.no/sisu/sisu_manual/sisu_help_sources/

SiSU home:
www.jus.uio.no/sisu/