

SA_MetaMatch: Document Discovery Through Document Metadata and Indexing

Christina Hiu-shan Yau
CS Department, University of Alabama
Box 87290
Tuscaloosa, AL 35487-0290
1-205-348-0997
hyau@cs.ua.edu

J. Scott Hawker
CS Department, University of Alabama
Box 870290
Tuscaloosa, AL 35487-0290
1-205-348-1667
hawker@cs.ua.edu

ABSTRACT

SA_MetaMatch, a component of the Standards Advisor (SA), is designed to find relevant documents through matching indices of metadata and document content. The elements in the metadata schema are mainly adopted from the Dublin Core (DC). The implementation of the XML metadata schema and coding follows the DC recommended guidelines. After metadata is generated manually for an unstructured document, or is extracted automatically from documents of well defined layout, they are stored in metadata files or in a repository. The indices of the descriptive metadata elements and that of the document content are generated. They are searched and compared to find related documents, based on our observation that if the metadata and high frequency index words of document content are related, then the corresponding documents are likely to be related as well. A ranked list of possible relevant documents is returned as the result. Several matching algorithms have been explored. We selected a sum of word-scored approach which not only gives relevant scores for the matched documents, but also gives an individual score for each of the matching words which provide hints for domain experts to grasp the concepts in the documents.

Categories and Subject Descriptors

I.7.5 [Document Capture]: Document analysis

H.3.3 [Information Search and Retrieval]: Information filtering, Query formulation, Search process, Selection process.

General Terms

Design, Experimentation, Verification.

Keywords

Dublin Core, metadata, index, document matching.

1 INTRODUCTION

1.1 NASA Technical Standards Program

The National Aeronautics and Space Administration (NASA) Technical Standards Program (NTSP) [8] provides access to technical standards documents and hosts a comprehensive

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference'04, Month 1–2, 2004, City, State, Country.
Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00

experience database of lessons learned, best practices and application notes. There are more than 1600 NASA agency-wide preferred technical standards¹, more than 45,000 standards from other government groups, and more than 95,000 standards from over 145 national and international SDOs (Standards Development Organizations), committees and working groups. The Lessons Learned and Best Practices (LLBP) include NASA-published lessons and links to over 30 lessons-learned databases from government and non-government organizations. To have the experience information readily available to the user of the technical standards, the Lessons Learned and Best Practices documents need to link with associated Technical Standards documents [4]. We have developed SA_MetaMatch, a tool to aid the discovery and linking of related standards and lessons learned.

1.2 Standards Advisor

SA_MetaMatch is one part of the NASA Standards Advisor. The Standards Advisor (SA) [5] enhances NTSP information retrieval by providing information portals to deliver documents specific to engineering activities, and by providing improved document search. Besides the search for documents that are relevant to user-entered keywords, it also searches for documents related to a given document. It matches lessons learned and other experience to relevant technical standards, and it builds hyperlinks to record discovered document relationships (after user confirmation), to facilitate maximum knowledge discovery. The tool that discovers and builds links between related standards and lessons learned is called SA_MetaMatch.

As a program to support standardization, the Standards Advisor follows applicable standards in its design. A widely adopted international metadata standard, the Dublin Core Metadata Element Set (DCMES) [1] and its recommended refinements [3], published by Dublin Core Metadata Initiative (DCMI) [2], is the basis of the data used for SA_MetaMatch. Applicable Dublin Core (DC) recommended metadata elements are selected and extended to form a document metadata model. The selected DC descriptive metadata elements and the whole document are processed with an indexer to generate a list of index words for matching.

1.3 DC Metadata and Indexing

Our approach to increase the productivity of the linking process is to combine the power of metadata and the power of indexing.

¹ Amount of NASA agency-wide preferred technical standards: ~1,000 standards from ~40 non-government standards developing organizations (SDOs), ~600 MIL (military) standards and specifications, ~60 NASA-published standards.

Both DC metadata elements and indexing have been used widely in information retrieval. Dublin Core elements, refinements and guidelines [1, 3, 11] are used for resource discovery in library systems. A document index is the main component which web search engines rely on. DC metadata elements return good search results, and controlled vocabularies, such as Library of Congress Subject Heading (LCSH), Medical Subject Heading (MESH), and other specific subject heading or thesaurus, are recommended to contribute to high precision. However, manual allocation of controlled keywords is costly and time consuming, often involving subject or domain knowledge. Also, the manually assigned words are subjective and thus metadata values can be inaccurate or incomplete about the concepts contained in a document.

Index with frequency of occurrence has the advantage of automatic generation using indexers and other commonly-available tools, providing an alternative, cost-effective solution to resource discovery. The important concepts of a document are likely to be found among the high frequency words. It is not unusual for technical specifications and standards to be over 100 pages, and glancing at an index may give a quick preliminary picture about the content of a document, thus helping the reader more quickly grasp the content, especially that of a long document.

However, the sole use of index and frequency can miss including some important concepts which appear in the document. This problem also exists in other metadata elements if controlled vocabularies are not consistently implemented. Besides reliability, another drawback of frequency of index word occurrence is its dependency on document size. The frequency of occurrence needs to be normalized by the document size.

In evaluating the relevancy of matched documents returned in a result list, the domain experts find a list of the index words ordered by relevance score or frequency of occurrence to be useful. The relevance score and frequency of occurrence needs to be normalized by the total score and document size respectively.

2 METADATA AND INDEX

2.1 Metadata Elements and Schema

The first version of the SA_MetaMatch metadata schema consists of twelve elements listed in Table 1. Ten of the fifteen DCMES recommended elements with their refinements [2, 3] are adopted from the DCMI recommendations. All the selected elements have the same name as that of their corresponding DC elements, but there are changes for some of the refinements. For the two remaining elements, audience is a DCMI but not DCMES recommended element, while Status is an element for existing standards document in NTSP. The SA_MetaMatch metadata are implemented in XML [13] format, referencing the DCMI published Guidelines [11].

2.2 Metadata Generation

Metadata can be extracted automatically from electronic documents with known structure. Lessons learned documents developed by NASA have a defined layout, and their metadata are extracted automatically by the SA_LLISMeta_AutoGen tool. However, for the documents with undefined or different structure, metadata generation needs manual effort. There are two options for capturing manually supplied metadata for electronic documents: annotation and template.

The first option is to add annotations into the document, such as XML tag values (for example, <Title> ... </Title> tags) marking the corresponding elements in the document. For this option, W3C tools such as Annotea Amaya [14] can be used to embed annotations based on the document metadata model. This tool uses XPointer to markup the annotated information. The semantically annotated document can be then processed to extract useful information, including validation against the schema.

Table 1. SA_MetaMatch metadata elements

SA_ MetaMatch Metadata Element	Attribute(s)	Element Refinement(s)	Element Encoding Scheme(s)
DCMES Element (* Non DC Element or Recommendation)			
Title	-	Alternative	-
Creator	Organization*	-	-
	Address*		
Subject	-	Classification*	NTSP Category*
		Controlled Keywords*	NTSP Keywords List*
		Other Keywords*	-
Description	-	Table Of Contents	-
		Abstract	
Publisher	-	-	-
Date	Date Type*	Created	DCMI Period ISO 8601 [W3C-DTF]
		Valid	
		Modified	
Format		Extent	-
		Medium	IMT
Identifier	ID Type*	-	URI / DocID*
Source	-	-	URI
Relation	Relation Type*	Is Version Of	URI
		Has Version	
		Is Replaced By	
		Replaces	
		Is Required By	
		Requires	
		Is Part Of	
		Has Part	
		Is Referenced By	
		References	
		Is Format Of	
		Has Format	
Other DCMI Recommended Element			
Audience	-	-	-
Additional Non DCMES Element			
Status	-	Validity Status Adoption Status	-

The second option is to create separate document elements that capture the metadata (for example, a Title object whose value is the title string). In this option, tool components are written that

interact with the user who selects a document element type and enters the values of that element (by cut and paste from the source document, for example). Form-based tools similar to the Nordic Dublin Core Metadata Creator [10] could help guide the user throughout the documentation process. Typically a form with fill-in boxes or pick lists is central to the tool. The tools can indicate which are the optional and which are the mandatory elements, and have on-line help. Some of them are built on the framework of a database which makes it easy to recycle portions of metadata which may repeat between data sets. With this option, the tools could be hard-coded for the specific document element types, or a model-aware tool could read the document model schema and interact with the domain expert to automatically create a template for the metadata elements that conform to the schema. SA_MetaMatch adopts the form-based method because standards and lessons learned documents are not available to edit for metadata tag insertion. A Metadata Generation Interface is developed for metadata generation and index generation. An extract of the metadata generated is shown in figure 1.

```
<?xml version="1.0" ?>
<Metadata>
  <title>Electrical Grounding Architecture For Unmanned
  Spacecraft</title>
  <subject>
    <controlledKeyword>Spacecraft</controlledKeyword>
    <uncontrolledKeyword>Grounding architecture, Design,
    Rationale</uncontrolledKeyword>
    <classification scheme="NASA PTSD">4000</classification>
  </subject>
  <description>
    <abstract>This handbook describes spacecraft grounding architecture
    options .at the system
    ...
    basis for understanding those choices and tradeoffs. </abstract>
    ...
  </description>
  <publisher>NASA</publisher>
  <date dateType="">February 17, 1998</date>
  <format>
    <extent>257K</extent>
    <medium>text/pdf</medium>
  </format>
  <identifier idType="docID">NASA-HDBK-4001</identifier>
  <relation relationType="hasPart">MIL-STD-1553</relation>
  <audience>spacecraft systems designers</audience>
</Metadata>
```

Figure 1. Extract of Metadata Generated

As Subject metadata with controlled keywords contribute to higher precision in search results, a domain-related dictionary and thesaurus are sources of reference. Keywords listed in existing lessons learned documents are selected from the LLBP controlled vocabulary list (~60 key phrases). WordNet [6] is a rich and popular online lexical database, and NASA also has developed its own online Thesaurus Machine Aided Indexing (MAI) program [9]. They provide related terms and thus are helpful for users when concepts have to be explored. WordNet returns a long list of terms, but many of the terms from this general database are domain irrelevant. The NASA thesaurus is more relevant, but it still returns some domain related but not relevant terms. SA_MetaMatch tries to build a small collection of thesaurus and abbreviation terms by extracting term definitions from the particular sections of standards documents. Links to WordNet, MAI, and the SA_MetaMatch collection are provided in the template to aid a user in selecting and filling controlled terms as Subject metadata. Domain ontology in the form of a thesaurus or other format would be useful in Subject assignment, too. Users

can browse to find the related broader or narrower term, that is, hypernym or hyponym for terms, to refine the search.

2.3 Index Generation

An index can be generated automatically with indexer tools. There exists a great deal of indexers nowadays because of Internet search. The indexers included in the open source list are: freeWAIS-sf, Harvest, Ht://Dig, Isite/Isearch, MPS, SWISH-e, WebGlimpse, and Yaz/Zebra [7]. After evaluating the list with the platform limitation and other criteria, SWISH-e [12] is selected as the indexer for SA_MetaMatch. SWISH-e is simple to install and use. It has a rich list of options which are configurable. The configuration file is used for both indexing and searching. It can filter stopwords, numbers, and words shorter than a certain length. It also supports stemming options, and can recognize HTML META tags. The ISAM (Index Sequential Access Method for Data Management which uses random access keys for index retrieval) index files can be searched with trace option to output an alphabetically-ordered list of index words, with their frequency of occurrence, word location(s) and other information.

Selected descriptive metadata (Title, Subject, Abstract) and the whole content of the document is processed with the SWISH-e indexer to generate an individual index for matching. To match only the important index words, which is likely to be of high frequency or in special metadata locations such as title, filtering is needed. The length of the document very often affects the absolute frequency of occurrence for a word in a document. To filter out low frequency words, the relative frequency is calculated and used. A Perl script program is used for the term extraction, frequency calculation and filtering.

Although some frequently occurring index words can be excluded from the resulting index by listing them as stopwords, there are still a great deal of domain related or document related stopwords that need to be filtered. The filtering of these words is especially important for indexing short documents. Examples of these additional words are as follows:

- common words and labels, including names (e.g. own organization name, own document ID, time, date, month names), section layout and label (e.g. "title", "chapter", "section", "topic", "keyword", "table", "figure", "page"),
- category related frequent words, e.g. "lesson learned", "recommendation", "technical specification", "standards", "report",
- user-specified words, customized by users; after the index words are sorted by frequency, the users can spot some topic-unrelated high frequency words, which are to be selected and picked out in a compiled list.

2.4 Metadata Storage

There are three options for the storage of metadata:

1. Embedded in the document;
2. In a separate standalone file that can be externally accessed and that points to the document;
3. In a database (repository) that holds the metadata and points to the document.

Metadata embedded in the document requires the capability to modify the document, and it makes searching and reasoning about metadata difficult, since it is intermingled with document content data (for example, by placing <META> tags inside the document). External metadata makes it easier for a computer

program to process the metadata. When the metadata is stored separately, association between the metadata and the primary content is needed. The separate file approach is easy to implement, but it lacks integrated management tools. To associate metadata with the content, resource references can be placed within the primary content so that format-savvy applications can make sure they have the right metadata. Alternatively, the string used in rdf:Description “about” attributes can be placed within the primary content. This allows software that understands the file format to verify that they have the right metadata. Use of a unique instance ID for the “about” attribute value can improve the reliability. The repository approach provides better data management but more technical involvement in implementation. A data repository can use a database to hold the fielded information in a table(s) or other data structure.

SA_MetaMatch stored metadata in standalone XML files in the early stage, and then an XML repository has been developed and is now used to store the data. With the repository, metadata can be managed better and manipulated more efficiently.

3 MATCHING DOCUMENTS

3.1 Matching Strategy

Two documents are related if they have matching terms found in title, subject, description, audience, or other criteria in the metadata, or have matching terms found in the high frequency words list generated from the metadata and document content. The relevance of each matching word depends on the location and the frequency of occurrences of the word in both the target and matching document. A matching document with high score of relevant metadata and index words is likely to be related to the target document.

A matching strategy and algorithm is needed to find relevant metadata and index terms and thus relevant documents. Weights are assigned to different metadata elements according to the importance of the elements, configurable by the user before running the matching program. Three algorithms have been derived and tested for the SA_MetaMatch program. They are:

- Comparison of index terms between same element,
- Comparison of index terms between alternative elements in addition to same element match,
- Matching word score and sum of word score for comparison.

The first simple matching algorithm (figure 2) just compares index terms only between same elements.

```
for each element in matching list {
    score += element_weight * %element_match (i.e.
    #match_word / #target_element_word)
}
E.g. percentage relevant score = [4 * (% match in title) + 3 * (%
match in subject) + 2 * (% match in scope / abstract) + 1 * (%
match in index)] / (4 + 3 + 2 + 1) * 100
for title weight = 4, subject weight = 3, abstract weight = 2, doc
index = 1
```

Figure 2. Matching Algorithm 1: Compare Index Terms Only Between Same Elements

The calculation for an improved matching algorithm (figure 3) compares index terms between alternative elements in addition to same element match, where the alternative element matched scored a discounted weight.

```
if (all words in a target element found matching in same comparing
element) {
    score += element_weight; (100% match)
    no need to compare with other element
}
else {
    %element_match = #match_word / #target_element_word;
    score += element_weight * %element_match;
    // compare the unmatched words with other elements starting from the
    closest weighted one until match word found in an element or
    none found for all elements;
    for each alternative match element alt_element, calculate the
    additional score add_score for matching words found in that
    element
    {
        wt_diff = abs(target_element_weight - match_element_weight);
        wt_diff_ratio = min((wt_diff / target_element_weight), 1);
        alt_element_weight = weight * (1 - wt_diff_ratio);
        %alt_element_match = #match_word_with_alt_element /
        #target_element_word;
        add_score += alt_element_weight * %alt_element_match;
    }
    score += add_score;
}
```

Figure 3. Matching Algorithm 2: Compare Index Terms Between Same and Alternative Elements

Another approach for the match is to calculate a score for each individual matching word and derive relevance from the sum of word scores (figure 4).

```
for each indexed element {
    // word refers to all words for target doc and
    // it refers to only words in the searched doc found to
    match that in target doc
    word score += word freq of that element *
    element_weight;
}
document score = sum of (word score);
relevance score of matching document = matching word score
/ (target word score + matching word score) * 100
```

Figure 4. Matching Algorithm 3: Calculate Word Score and Compare Sum of Document Word Scores

For the word score matching approach, the word list of the target document has all the terms listed and a score calculated for each word according to the frequency and location (i.e. type of element) of the terms (figure 5). Each frequency of occurrence for the term in document metadata and content scores according to the weight assigned to the location. The words in the list are then sorted by the resulting score in descending order and the low-scoring ones are removed according to the filter out percentage specified in the configuration file.

The total relevance score for all three approaches are the normalized sum of matching index words relevance score. SA_MetaMatch returns a list of possibly related documents sorted by the normalized scores (figure 6). The result interface links the original documents, metadata, word score and other relevant information to the target and matched document in the result list. The users can access those relevant data by just a click (figure 7).

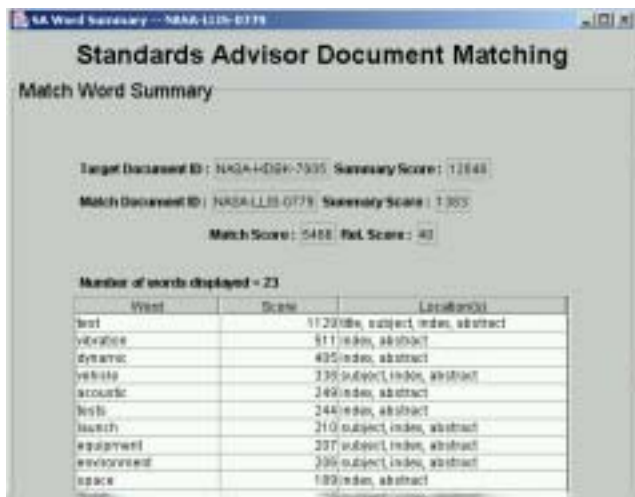


Figure 5. SA_MetaMatch Word Summary Interface Screen



Figure 6. SA_MetaMatch Result Interface Screen

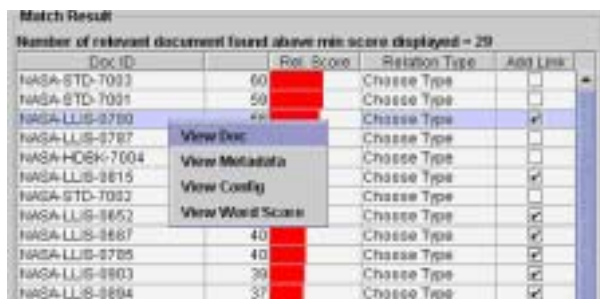


Figure 7. SA Result View Popup Menu

The scores of the matching documents and words give a hint to their importance and the relevance of the terms in the document. This can help the domain experts to quickly identify the main concepts in the document without reading the document.

3.2 Other Strategies

In addition to matching index words from document metadata and document content, other strategies may be employed, such as

- Recognize special pattern words, such as NASA-STD-7002, which indicates a relation with another standard document, according to the naming conventions of Standards Development Organizations,
- Remove irrelevant words selectively,
- Usage of a word in parts of speech (verbs, nouns, ...),
- Expand or refine search by applying semantically related terms in an ontology, including synonyms in a thesaurus, or broader or narrower terms in a concept hierarchy (taxonomy),
- Apply other ontology relationship including homonym, antonym, etc.,
- Apply more ontology (domain model) processing of related terms and constraints.

The special patterns recognizing and selective words removal has been implemented and tested in the SA_MetaMatch prototype.

4 RESULTS

Two sets of sample documents, including 10 NASA generated technical standards and 70 lessons learned (LL) documents were used to test SA_MetaMatch. These included documents identified by domain experts to be related and documents with undiscovered relationships. The results of SA_MetaMatch are verified by the domain experts.

The relevant document ranks utilizing the second and the third algorithm (figure 3 and 4) were found to be close in performance, and they are generally better than that for the first algorithm. Comparing the second and the third algorithm, the third one is preferred because it provides an individual word score which helps a domain expert assess each document's relevancy. Good performance was found when the minimum relative frequency for index words was set at about 0.2 and the low score word filtering out percentage was set at 25-50% depending on the length of the document. The ratio of title : subject : abstracts : index $\sim$ 50:40:30:1 for the third algorithm (figure 4) gave good results because it is based on frequency of occurrence of index words as well as their locations in metadata elements or document content.

SA_MetaMatch has been helpful for NASA's domain experts in finding relevant documents. It is found that on average about 70% of the known relevant documents rank in the top 25th percentile and about 90% of the known relevant documents rank on the top 50th percentile in both configurations of word filtering.

The additional strategies applied were also found to be useful in finding relevant documents. Special pattern recognition successfully discovers all reference documents cited (see figure 6). Selectively remove the irrelevant words from the indices, experiments show that the average amount of known relevant documents ranking in the top 25th percentile has increased to about 80%, and that in the top 50th percentile has increased to about 95%.

5 CONCLUSION

This study found that Dublin Core recommendations and guidelines provided very useful references for the XML metadata schema modeling and construction. The indices of the selected descriptive document metadata and the document content were found to be useful for identifying documents related to each other when used with a scoring algorithm based on frequency of index word occurrence as well as their locations in metadata and document content. Experiments have guided us to settings of

element weights and filtering values that provide good match results.

In calculating the document relevancy score, the indices for both the metadata elements as well as for the document content were found to be useful components. The individual word score calculated by combining the matched index words from metadata and document content has shown to facilitate user understanding of the concepts in standards and lessons learned documents, and the derived algorithms have shown a positive result in assisting domain experts in relevant document discovery.

The SA_MetaMatch prototype has proven to be useful to NASA in discovering related standards and lessons learned documents. It has improved the productivity of the document matching and linking process by automatically identifying lessons learned and standards documents that are probably related to each other. The software application has assisted the domain experts to find new links previously undiscovered by manual examination.

6 FURTHER STUDY

More research and study on the strategies to further improve the result of relevant document identification are necessary. The following are suggested:

- Apply stemming
Word stemming algorithms are to be studied and employed. The current SA_MetaMatch implementation of index words does not include the word stemming option in the SWISH-e indexer. The reason is the many errors found in the Porter Stemmer algorithm. New word stemming algorithms and improvements on the most popularly employed Porter Stemmer algorithm should be helpful.
- Implement word phrase index term
Word phrase matching for index terms should give better results by solving the high-scored matching index words found to be irrelevant in context.
- Analyze usage of a word and natural language processing
Usage of a word in parts of speech (verbs, nouns, adjective, or others), and further usage of a word in heading title, figure caption, etc, are to be analyzed. These can be important components for the relevancy calculation.
- Apply related terms and constraints
Expand or refine index terms by applying semantically related terms in domain ontology, including synonyms in a thesaurus, or broader or narrower terms in a concept hierarchy (taxonomy). A thesaurus can be referenced to find equivalent or similar terms, as well as broader or narrower terms to refine search.
- Apply other ontology relationships
Other ontology relationships including homonym, antonym, etc. can be applied to distinguish words of different context and to eliminate unrelated documents.
- Adopt future metadata and ontology language

XML has been employed as the metadata implementation language, but in the future RDF, DAML+OIL, OWL or another ontology language should be adopted when the application of the language has matured with enough supporting tools in the market.

- Learn from query
Related words can be learned from query. Search query feedback regarding whether documents selected are related is helpful in identifying related terms. Related terms can be catalogued and archived for future use.

7 ACKNOWLEDGMENTS

The authors wish to acknowledge the support for this work by the National Aeronautics and Space Administration (NASA) Marshall Space Flight Center in Huntsville, Alabama under NSSTC Cooperative Agreement NCC8-256

8 REFERENCES

- [1] DCMI. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*, 2 June 2003.
<http://dublincore.org/documents/dces/>
- [2] DCMI. *Dublin Core Metadata Initiative*. <http://dublincore.org/>
- [3] DCMI Usage Board. *DCMI Metadata Terms*, 4 March 2003.
<http://www.dublincore.org/documents/dcmi-terms/>
- [4] Gill, P., Vaughan W., and Garcia D. *Lessons Learned and Technical Standards: A Logical Marriage*. *ASTM (American Society for Testing and Materials) Standardization News*, November 2001.
http://standards.msfc.nasa.gov/LL_TechStdLogicalMarriage.pdf
- [5] Hawker, S. *The Standards Advisor Project*, University of Alabama. <http://www.cs.ua.edu/sel/standardsadvisor/>
- [6] Miller, G., Fellbaum, C., Tengi, R. *WordNet: A Lexical Database for the English Language*.
<http://www.cogsci.princeton.edu/~wn/>
- [7] Morgan, E. L. *Comparing Open Source Indexers*. Infomotions, Inc., 28 May 2001.
<http://www.infomotions.com/musings/opensource-indexers/>
- [8] NASA. *NASA Technical Standards Program*.
<http://standards.nasa.gov/>
- [9] NASA Scientific and Technical Information (STI) Program. *NASA Thesaurus Machine Aided Indexing*.
<http://mai.larc.nasa.gov>
- [10] Nordic Metadata Project. *Dublin Core Metadata Creator*.
<http://www.lub.lu.se/cgi-bin/nmdc.pl>
- [11] Powell A., Johnston, P. DCMI. *Guidelines for implementing Dublin Core in XML*, 2 April 2003.
<http://www.dublincore.org/documents/dc-xml-guidelines/>
- [12] SWISH-e. <http://www.swish-e.com>
- [13] W3C. *XML Schema*. <http://www.w3.org/XML/Schema>
- [14] W3C. *Annotea Project*. <http://www.w3.org/2001/Annotea/>
Amaya Home Page. <http://www.w3.org/Amaya/>