



# LIRICS

## Deliverable <D1.3>

### <Quality Assessment>

Project reference number	e-Content-22236-LIRICS
Project acronym	LIRICS
Project full title	Linguistic Infrastructure for Interoperable Resource and Systems
Project contact point	Laurent Romary, INRIA-Loria 615, rue du jardin botanique BP101. 54602 Villers lès Nancy (France) romary@loria.fr
Project web site	<a href="http://lirics.loria.fr">http://lirics.loria.fr</a>
EC project officer	Erwin Valentini
Document title	Quality Assessment
Deliverable ID	<b>D1.3</b>
Document type	Report
Dissemination level	Public
Contractual date of delivery	<milestone>
Actual date of delivery	28 February 2007
Status & version	Draft
Work package, task & deliverable responsible	WP1, Unis
Author(s) & affiliation(s)	Lee Gillam (UniS), Neil Newbold (Unis)
Additional contributor(s)	<optional contributor(s)>
Keywords	ISO, Terminology, Quality Assurance, Readability

### Document evolution

version	date	version	date
1.0	02/03/07		

# 1 Introduction

The objective of work package 1 is to guarantee that the documents, test suites and APIs produced within the project are designed in accordance with the general principles and structure of international standards within ISO, and, as they reach maturity that there are means and measures to check technical soundness and adequacy with the market.

The task of Quality assessment (1.3) is intended to introduce specific QA steps that enables ease of validation of documents by the eventual markets, and more immediately by the experts and implementers from industry through the Industrial Advisory group and by the ISO member bodies. In the longer term, this could substantially reduce time to market of standards; in the more immediate term, standards authors will need to assess the feedback produced and determine the appropriate course of action. ISO member body commentary has, in some ways, fed the requirements for the approach taken.

Writing standards requires a specific approach to style and vocabulary. International standards have specified structures, and content control approaching controlled authoring would be valuable in this arena to ensure that, for example, definitions provided can follow the principle of substitutability (whereby they can be used almost directly in place of words in the text and other definitions). The ability to undertake such a (semantic) task requires the document to be relatively well-written, with consistent terminological use and removal where possible of verbiage and ambiguity.

To provide for a variety of aspects of additional quality control, in addition to the extant processes of ISO, we have evaluated the efficacy of the University of Surrey Department of Computing's content analysis applications (System Quirk), developed in prior research, including EU-co funded projects. Various components have proven to be suitable for assisting reader and writer alike. The work covers the integration and use of supporting resources and components for the standards development process, including a Plain English thesaurus, lookup of ISO TC 37 terminology provided from a terminology management system (TMS) via ISO 16642, automatic terminology discovery using statistical and linguistic techniques, and readability metrics.

Efforts have been undertaken to integrate these components within an existing framework to demonstrate the potential for controlled authoring based on some of the very standards being used and produced within LIRICS – in a sense, this demonstrate LIRICS “eating its own dog food” or “drinking its own champagne”, depending on national preference. The result of these efforts leads us to the development of an assistive tool for authors of standards based around LIRICS work.

Initial experiments helped us to provide some additional commentary into ISO on a few standards documents at various stages of the ISO process; fuller sets of commentary for the LIRICS standards are currently in production and this deliverable presents some examples of how these are formulated. Human interpretation of, and action upon, the results being produced by these components is still required to varying extents, however the analysis of language simplicity and consistency, identification of known and unknown terms, and the generation of “understandability” metrics have all been trialled and demonstrate interesting and potentially highly-valuable results.

The LIRICS proposal considered that, longer term, a content management system for developing standards was envisaged, but that this was beyond the scope of LIRICS. These efforts are a major step towards the provision of such a system although the means of dealing with overlaps and inter-dependencies amongst the various components, and of providing further improved analysis, will require further treatment.

## 2 Background

### 2.1 Readability

A measure of the *accessibility* of written text is referred to as its readability. It is a “quality of a written or printed communication that make it easy for any given class of persons to understand its meaning, or that induces them to continue reading” (English and English 1958) or, more simply, the “ease of reading words and sentences” (Hargis et. al. 1998). A variety of measures of readability have been constructed on the basis that sentence length and word length, in some cases as a function of the number of syllables, are determining factors (Kitson 1921).

The most common readability measurement is the *Kincaid Formula* (Kincaid et al 1975). Other readability measures are the *Flesch Index* or Flesch Easy Reading Formula (Flesch 1948), the Fog Index (Gunning 1952), the Simple Measure of Gobbledygook (SMOG, McLaughlin 1969), and the Automated Readability Index (ARI, Smith and Senter 1967). The results of applying the various formulae proposed attempt to indicate the level of education or reading level, or provide a difficulty score on a scale of 1-100. These techniques rely on calculations based on counts of certain features of the text. The features used are presented in Table 1. It should be noted that computer programs can count the number of characters more accurately than the number of syllables since there are debates over what is counted as a syllable.

	Kincaid	Flesch	Fog	SMOG	ARI
Sentence length	✓	✓	✓	✓	✓
Characters/word					✓
Syllables/word	✓	✓			
Complex words count (more than three syllables)			✓	✓	
Scale	Grade level	0-100	Grade level	Grade level	Grade level
Ideal outcome	7-8 (13-14)	100	7-8	7-8	7-8

**Table 1: Features of readability metrics**

Readability measurements have proven popular because, in principle, they enable objective determination of the effective audience for a text. Educators use readability assessments to help select the appropriate reading material for their students, and the measurements identified above have been variously used by health authorities such as the Veteran's Association and agencies of the US government, including the Department of Defence and Navy, in tasks such as the assessment of insurance forms.

### 2.2 Plain and Simplified English

The Plain English Campaign (PEC) was formed in 1979 by Chrissie Maher OBE to help organisations produce documents – especially those forms, leaflets, agreements and contracts that are meant to be understood by the public at large – that everyone can read,

understand and act upon. PEC provides rules and guides that can help make writing clearer by avoiding verbose sentences, jargon and other confusing uses of language, for example:

- Short sentences are preferable, with an average length of 15–20 words
- Language should be positive and active verbs should be used wherever possible to help make documents sound fresh and professional
- Documents should contain 80–90% active verbs
- Use the simplest word that fits
- Avoid words and phrases with multiple meanings
- Avoid using slang and jargon
- Substitute unnecessary complex words with simpler alternatives.

PEC provides a guide suggesting hundreds of plain English substitutions for verbose words and phrases. For example, “in accordance with” can be replaced with one of “as under”, “in line with” or “because of”. However, PEC advise that many of these alternatives won't work in every situation.

In addition to PEC, the European Association of Aerospace Industries (formerly AECMA, now ASD) has a specification for authoring of aircraft documentation. English is the international language in the aerospace industry, but often not the native language of its readers. Many of these readers can be confused by complex sentence structures, polysemy and synonymy. *ASD Simplified Technical English* (STE), formerly AECMA Simplified English, was developed to control the language and writing style of English-language documentation to help such users. Studies have shown that the benefits to native English speakers are also significant, and include reduced error rates and a decrease in the time taken to complete a task.

STE is a reduced form of English aimed at removing the ambiguity of complex statements. STE is a controlled language which uses limitations on grammar and style and a restricted base vocabulary of 1,000 words. Writers are only allowed to use words from the controlled dictionary, where each word has been chosen for simplicity and ease of recognition and has one agreed definition. The controlled dictionary has sufficient words to express any (aerospace) technical sentence. STE contains around 60 writing rules; similar to PEC, the specification recommends avoiding the passive voice, being specific, and limiting verbiage. More specific advice is given: the recommended maximum for a procedural sentence is 20 words, and for a descriptive sentence the limit is 25 words. In addition, the rules advise against using clusters of more than three nouns: *runway light connection resistance calibration* should be written as *calibration of the resistance on a runway light connection*. PEC and STE should both increase reading speed and ease translation.

### 2.3 Terminology

Terminology can be considered as a field of science concerned with concepts and terms, in one or more languages, of specialisms. The results of terminology work are made available to users in the form of terminology collections: lists of specialised terms, glossaries or technical dictionaries. The language of specialisms, and of science, can be difficult to understand due to the profusion of terms contained within the texts that may be of the specialism or seek to explain the specialism. Over-extensive use of complex terms, pejoratively referred to as jargon, can exclude and alienate readers. The suggestion is that technical terms are difficult and unnecessary, and that the same meaning could have been conveyed using simpler everyday language.

ISO TC37 has been an important committee of ISO for 50 years and has added to work carried out by Eugen Wüster (1898-1977), regarded as one of the founders of terminology as a scientific discipline. Subcommittees of ISO TC37 are concerned with various issues relating terminology including presentation, structuring, related metadata, and the management of systems and interoperability between systems. ISO TC37 standards set rules for the provision of terminology within standards: in the UK, the British Standards Institution (BSI) has standards for standard, BS 0, that normatively reference TC 37's ISO 10241: International terminology standards – Preparation and layout. The terminology of terminology, and of associated practices, is incorporated within TC37's documents, and the subject of almost continuous review.

In recent years, ISO TC37 have produced an ISO standard for a terminological markup framework (ISO 16642) that enables interoperability between terminology collections via comparisons of the metadata descriptors (data categories – ISO 12620) used within them. This provides a further step towards the computability of terminology, and a means by which extant terminologies can be used within other systems. The task of terminology acquisition and structuring, a precursor to provision in such a computable format, has been variously considered from linguistic and statistical perspectives. An overview of various techniques and systems has been provided, and a hybrid method has been explored that employs both statistical and linguistic techniques has been discussed (Gillam, Tariq and Ahmad 2007; Gillam and Ahmad 2005).

Producing a useful and consistent terminology collection is not without its challenges. A concept can have differing designations and definitions amongst different communities with little consensus - denominative variation; one specialist can name a concept in different ways - self-variation; different specialists can express the same idea in different ways - hetero-variation. Terminological variation can appear among different authors for various reasons, influenced by geography, chronology or social factors. These variations can become apparent in the documents produced, especially in multiply-authored situations, but only following sufficient analysis of individual documents (self-variation) or a collection of inter-related documents (hetero-variation / denominative variation). Lack of consistency is likely to confuse a reader, particularly a reader of a standard. Traditionally, however, the burden of determining terminological consistency or variation has been placed firmly upon the reader; this burden has been greater since the standard may contain terms defined in a document not in the reader's possession and which can only be obtained at an additional cost.

## 2.4 Integration

Plain English and Simplified English would appear to provide for rules of how readability can be improved, and the expectation would be that the application of such techniques as using simplification lists would produce a better readability score. Avoidance of "jargon" would appear to be substitutable for ensuring that complex terminological constructions within the document are well defined and that these definitions are easily found and substitutable at the places in the document in which they are used. The various readability metrics appear to take little account of any difference between known and defined terms and those which are not made so accessible. Additionally, multiword expressions and proper nouns are not counted as complex words, and not all multisyllabic words are difficult to understand. For example, "spontaneous" is generally not considered a difficult word, despite having four syllables. Since the majority of readability metrics seem to ignore such factors, with the exception of counting the number of "complex words" in the Fog Index, a new kind of readability metric may be needed.

The ASD notion of a controlled dictionary, where one word has one definition, provides a nice parallel to terminology work, suggesting that providing for management of terminology within such a system has substantial importance. This includes facilities both for dealing with "known" and "unknown" (discovered) terms. Improving the consistency with which terminology is used across (standards) documents by making it available to authors and readers appears to lead us in the direction of ASD. The increasing complexity of technical content and the ever-greater demand for accurate information and understanding suggests a move away from *ad hoc* approaches. Even though ASD Simplified Technical English was specifically designed for use with documentation in the aerospace sector, its principles may be applicable to other areas, particularly standards.

Combining the resources identified earlier in this section is not a straightforward task. Overlapping the identification of known and unknown terms, and of two means of deriving these unknown terms, necessitates efforts in treating the overlaps in the results. The list of PEC simplifications entails some suggestions that may overlap with terminology or vice-versa. Furthermore, attempting to automate the discovery of those situations in which PEC simplifications are valid requires treatment of a number of contexts in which these constructions exist in order to generate the appropriate rules and exceptions.

One of the most common criticisms levelled at ASD and PEC is that science depends on scientific language and, likewise, technology depends on technical language. Science and technology cannot easily be separated from how they are written, and human authors cannot easily be separated from packing semantics into the most economical terminological shape, including the use of abbreviated forms. It would be difficult to present scientific and technical knowledge entirely in common wordings. However, scientific writing can be made more difficult than it need be, and authors can form a habit of locking themselves and their peers into unnecessarily complicated construction. One can only hope that authors are not consciously seeking to distinguish their discourse as that of the intellectual elite, where the pejorative use of jargon would be apt. It is, therefore, on the basis of accidental complexity that we consider these efforts.

### **3 Document Content Management System**

To automate quality assurance, a prototype document content management system was considered. The purpose of such a system is to examine a text and identify existing and possible terms, measure its readability and offer plain English alternatives for verbose words and phrases and to provide meaningful feedback to the document authors. Occurrences of existing terms in a text would be identified using a terminology that would be integrated via ISO standards 16642 and 12620. The terms and definitions would be available as sources of extra information to help clarify the meaning of the known term and to help understand its use in running text. If ISO definitions are substitutable for the term in a text then this can be explored by linking a term with its definition, though the result may entail adaptation of either the text or the definition. The purpose is to allow the relationship between a term and its definition to become clearer and more accessible.

We formulated, also, a hybrid of two separate methods for recognising potential new terms in a text. The linguistic method uses combinations of parts of speech to identify multiword expressions based on patterns variously described in the literature. A multiword noun that occurs more often than a particular frequency threshold will be considered as a potential new term. The second approach identifies the keywords in the text by determining the thresholds of z-scores of frequencies and weirdness of each word in the text, and applies collocation statistics, following Smadja, to words surrounding these keywords (Gillam 2004) on the basis of which a terminological structure could be induced. These words in conjunction with the keyword they relate to can then be considered as a possible additional terminology. In each case, the focus is on relatively frequent occurrences since these provide some degree of coherence to the document.

The prototype document content management system will perform readability metrics on the text to give an indication how well written the document is. To help improve the readability of the document suggested improvements for writing from ASD Simplified Technical English and the Plain English Campaign will be incorporated. Suggested replacements for commonly used verbose and difficult phrases will be highlighted. These replacements can then be substituted for the original text to help improve the readability of the document.

#### **3.1 Implementation**

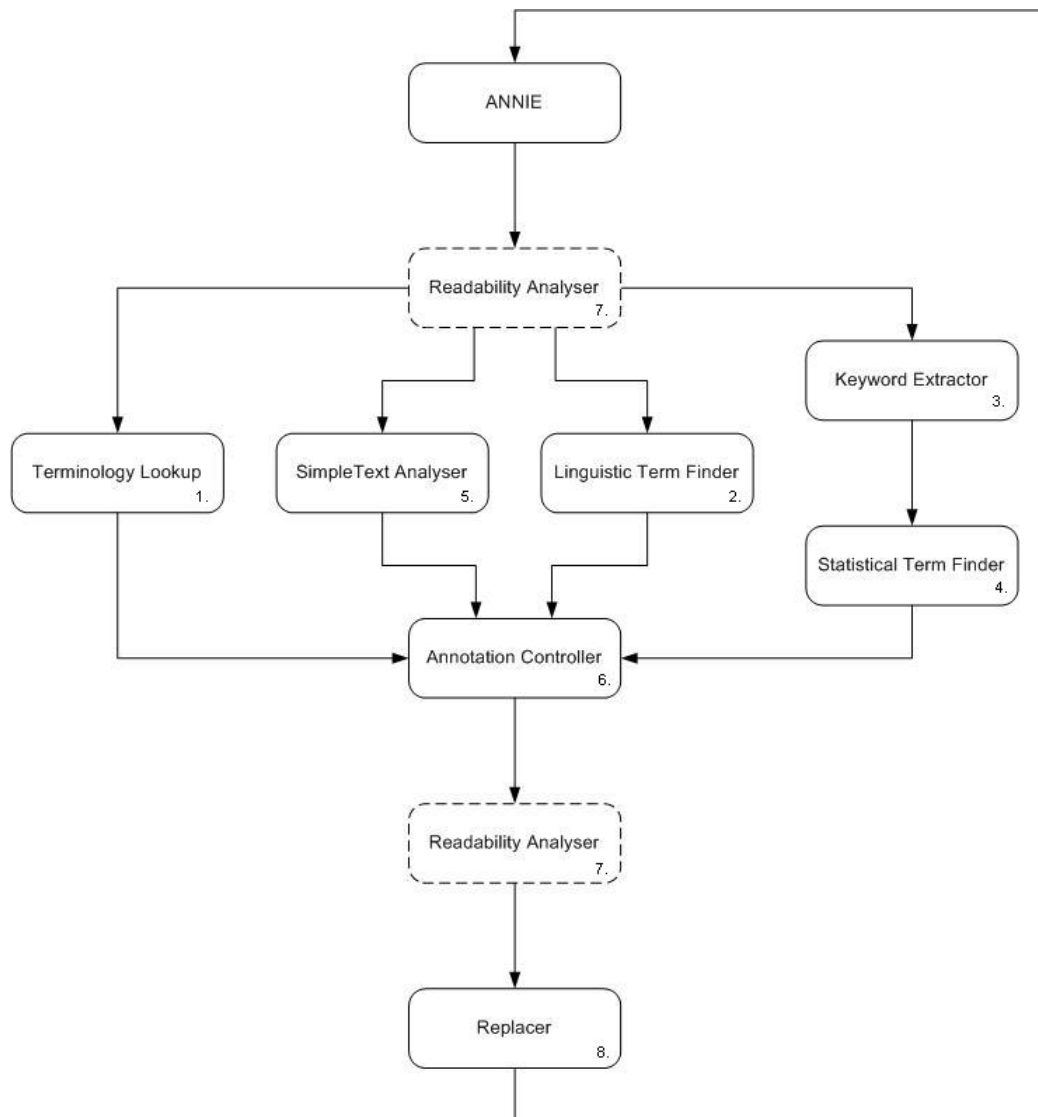
The University of Sheffield's General Architecture for Text Engineering (GATE, Cunningham et al 2002) was selected as a basis and front-end for the implementation. GATE is established within the NLP community and has been used for numerous research projects. The GATE interface allows for different "processing resources" to be executed in sequences in what is referred to as a pipeline; the user can order the execute of these processing resources. A set of reusable processing resources for common NLP tasks is provided with GATE, packaged together to form A Nearly-New Information Extraction (ANNIE) system.

We decided to expand the function of GATE into a system capable of document content management. We utilised the existing plug-ins from ANNIE, the tokeniser, sentence splitter and POS tagger, for the preliminary tasks. We then developed, incrementally, a set of new

processing resources that adopted techniques for improving the readability of documents by incorporating the use of terminologies, readability measurement and suggested improvements for writing from ASD Simplified Technical English and the Plain English Campaign. Eight new processing resources were developed which are detailed below:

1. Terminology Lookup
2. Linguistic Term Finder
3. Keyword Extractor
4. Statistical Term Finder
5. SimpleText Analyser
6. Annotation Controller
7. Readability Analyser
8. Replacer

The pipeline for these resources is shown in Figure 1, below, with brief descriptions of each component following. It should be noted that the Readability Analyser can be run at two separate points in the pipeline.

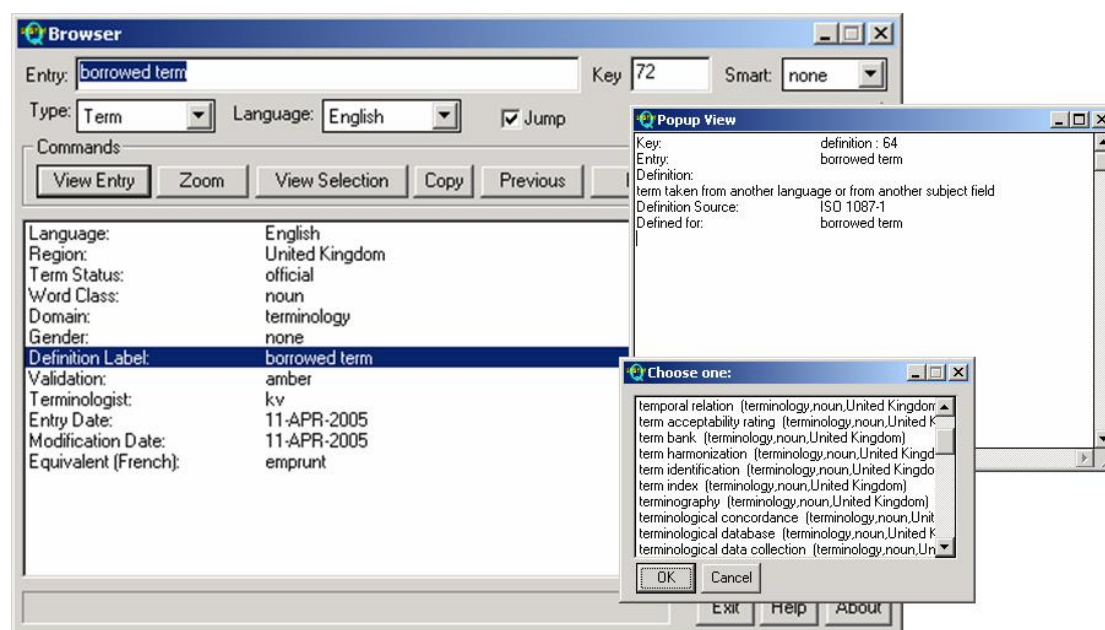


**Figure 1: Pipeline for the prototype document content management system**

### 3.1.1 Terminology Lookup

This resource is run after the ANNIE tokeniser and sentence splitter. It uses an XML file that comprises all the terms and definitions for the subject field. The XML file is a small-scale export of a terminology database in the meta-model format specified in the ISO 16642

standard. This terminology was collected from existing ISO standards during the project and inserted into a terminology management system (System Quirk's Browser/Refiner application – Figure 2). The database contains the terms from 15 standards documents, in English and French, as shown in Table 2 below.



**Figure 2: An example of a term, with its ID and definition.**

The terminology database contains 860 defined terms of which 648 are in English and 212 are French. The number of terms used from each standard is detailed in figure 3.

Standard	Term Count
ISO 639-1	3
ISO 639-2	5
ISO 860	1
ISO 1087	3
ISO 1087-1	180
ISO 1087-2	229
ISO 2022	16
ISO 2382-4	46
ISO 3166-1	2
ISO 3166-2	1
ISO 4873	3
ISO 8601	17
ISO 8879	293
ISO 12615	15
ISO/CD 24610-1	21

**Table 2: Number of terms taken from ISO standards currently contained in the terminology database**

The input parameters for the Terminology Lookup processing resource are the location of the ISO 16642 conformant XML file containing the full terminology and a language identifier used to identify the subset of the terminology collection to be used. The terminology lookup plug-in examines a document and annotates those term entries found in the text. The annotations created by this plug-in are called 'KnownTerm'. These annotations store the ID and the definition of their terms as well as a substitute option. This option allows the user to replace a

single occurrence of a term within the text with its ISO specified definition. According to ISO specifications, any terminology definitions should be replaceable for the term within a document without altering the meaning. The default setting for the substitute option is 'no' but if it can be altered to 'yes' to signify that this term should be replaced by its definition in the running text. The actual process of substituting the text is performed at a later stage by the Replacer (3.1.8) plug-in.

### 3.1.2 Linguistic Term Finder

The *Linguistic Term Finder* is run after the ANNIE tokeniser and sentence splitter. This plug-in determines all the compound nouns in the document according to specified patterns of part of speech annotations (e.g. in Jacquemin 2001). Multiword (noun) expressions which occur in the document with a greater frequency than the input parameter for the frequency threshold are annotated as 'LinguisticTerm'.

### 3.1.3 Keyword Extractor

The *Keyword Extractor* is run after the ANNIE tokeniser and sentence splitter. This plug-in calculates the frequency and weirdness of individual words, as defined by Gillam (2004). The user supplies a file of words and their frequencies in a reference corpus: we use frequency information from the 100 million word tokens of the British National Corpus (BNC). Frequently used words in the document which have a low frequency in the BNC, i.e. can be called "weird", are annotated as 'Keyword'. These annotations can be adjusted by the user using the input parameters for frequency and weirdness z-score thresholds. The extracted keywords may, in some cases, already be annotated as "KnownTerm".

### 3.1.4 Statistical Term Finder

The *Statistical Term Finder* is run following the Keyword Extractor (3.1.3). This plug-in examines the neighbouring words around a keyword and identifies recurring patterns. Input parameters include neighbourhood size (distance from this keyword) and weirdness threshold for inclusion. If a word consistently appears in the user-defined neighbourhood size with a predetermined level of weirdness then it is considered a potential new term. These new terms are annotated as 'StatisticalTerm'.

### 3.1.5 SimpleText Analyser

The *SimpleText Analyser* is run after the ANNIE tokeniser and sentence splitter. This plug-in uses a dictionary of words and phrases identified as verbose by either the Plain English Campaign or ASD Simplified Technical English. This Plain English Campaign information can be downloaded from <http://www.plainenglish.co.uk/alternative.pdf> and the ASD-STE100 Specification can be requested from [http://www.simplifiedenglish-aecma.org/Simplified\\_English.htm](http://www.simplifiedenglish-aecma.org/Simplified_English.htm). The dictionary contains a list of 1302 phrases and offers a selection of one or more alternatives for each. The SimpleText Analyser identifies these phrases within the text and produces 'SimpleText' annotations. The annotations contain information regarding the potential replacements for the expression. Up to five replacements can be stored within the annotation. The annotation also contains a "best replacement" feature which can be amended by the user to contain the replacement they consider the most suitable. If this feature is left blank, it signifies that none of the suggested replacements are suitable. This feature is used at a later stage by the Replacer (3.1.8) processing resource.

### 3.1.6 Annotation Controller

The *Annotation Controller* runs after the terminology lookup (3.1.1), linguistic term finder (3.1.2) and statistical term finder (3.1.4) plug-ins. The annotation controller combines the 'StatisticalTerm' and 'LinguisticTerm' annotations to create a 'DiscoveredTerm' annotation. These annotations contain information detailing whether the 'DiscoveredTerm' originated from linguistic or statistical analysis. This information is shown as 'ValidLinguistic' and

'ValidStatistical' accordingly. Any discovered term which is both linguistically and statistically valid is a strong case for addition to the terminology. The Annotation Controller is used to reduce the quantity of overlapping annotations being produced by prioritising some annotations over others. If a term is already annotated as a 'KnownTerm', that annotation takes priority over the 'DiscoveredTerm' annotation, which is excluded from consideration. A further consideration is that a 'DiscoveredTerm' may contain an 'KnownTerm', in which case the 'DiscoveredTerm' is retained. The annotation controller prioritises 'SimpleText' annotations lowest, removing those 'SimpleText' annotations which overlap annotations for 'KnownTerm' and 'DiscoveredTerm'. This feature was added after discovering many suggested replacements were redundant as they were contained parts of the terminology.

### 3.1.7 Readability Analyser

The *Readability Analyser* can be run over the whole text after the tokeniser and sentence splitter. The readability analyser takes no input parameters and produces two annotations for users to examine. These annotations are called 'Count' and 'Readability' and cover the whole text of a document. The 'Count' annotation stores the number of words, syllables, sentences, characters and polysyllabic words contained within a document. A polysyllabic word is defined as one containing three or more syllables. The number of polysyllabic words in a document is used for the calculation of certain readability formulas. The 'Readability' annotation stores the results of the various readability measurement formulas performed by the 'Readability Analyser'. The formulae that are applied to the text are the Kincaid formula, Flesch Index, Fog Index, SMOG and ARI readability measurements. The readability analyser was devised so that any words previously annotated as terminology are not considered as complex words when calculating the Fog Index and SMOG formulas. This means that the readability scores of a document can be improved by adding terms to the terminology, emphasising the need to manage terminology. These new terms should be added with definitions in accordance with ISO specifications. The Readability Analyser can be run at two different points in the pipeline. This allows for comparison of the results of the readability formula with and without terminology annotations. The effect of adding words to the terminology can then be clearly observed. The analyser allows numerous readability annotations to be created and by preserving the scores over iterations of the document, the historic readability of the document can be values can be examined.

### 3.1.8 Replacer

The final processing resource is the *Replacer* which is run after the 'SimpleText Analyser' and the 'Terminology Lookup'. The replacer plug-in automatically substitutes the text in a document with user-selected "best replacements" (3.1.5). If no best replacement is selected then the text is left unchanged. Also, text annotated as a 'KnownTerm' can be replaced with its ISO definition. This replacement only occurs if the substitute option in the Terminology annotation is set to 'yes' (3.1.1). In the process of replacing text, the replacer removes most of the existing annotations on a document. The only annotation left on a document by the replacer is the 'Readability' (3.1.7) annotation. This annotation is kept so that the user can compare the results of the readability analyser to previous executions. The replacer also creates a new annotation called 'ReplacedText' which covers the start and end point of the text that was replaced. This annotation stores the word or phrase which was originally in the document before the replacer was executed.

Once the replacer has finished, the whole procedure can be repeated with the ANNIE plug-ins being executed again to re-tokenise and sentence split the document. The readability analyser can be run again to show the effect that the replacements had on the readability scores of the document. The user can then decide whether any further replacements or additions to the subject terminology (via the terminology XML file) are appropriate.

## 4 Automated Quality Assurance

To demonstrate results of this analysis for the purposes of this deliverable, two standards being developed within the LIRICS project, and at various stages of the ISO process, have been analysed using this prototype system. The documents 'Lexical markup framework (LMF)' (at Draft International Standard stage<sup>1</sup>) and 'Syntactic Annotation Framework (SynAF)' (at Working Draft stage<sup>2</sup>) were chosen to show the output obtained from the various stages of the analysis.

### 4.1 Terminology Lookup

All known terms were annotated including those occurring with another term within the terminology annotation. For example, the known term 'object language' had another known term 'object' annotated within it. The annotation allows access to the definition for the term. An example of how the terms were annotated in the document 'LMF' is shown in figure 4.

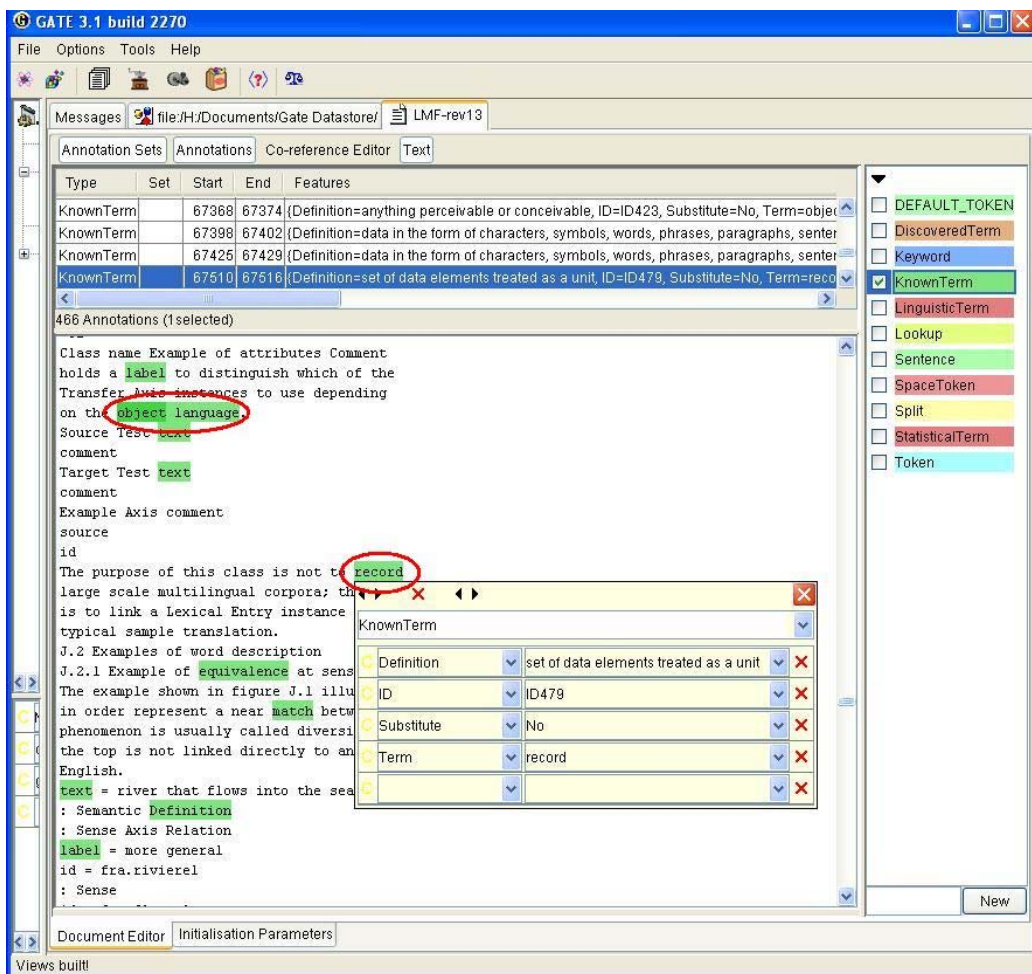


Figure 4: A screenshot of the 'LMF' document in GATE displaying the terminology annotations and a popup window of annotation information for the term 'Record'.

<sup>1</sup> Revision 13, available at: [http://lirics.loria.fr/doc\\_pub/N330\\_LMF\\_rev13\\_For\\_CD\\_Ballot.pdf](http://lirics.loria.fr/doc_pub/N330_LMF_rev13_For_CD_Ballot.pdf)

<sup>2</sup> Revision X, available at: [http://lirics.loria.fr/doc\\_pub/SynAF\\_WD\\_2006-01-22.pdf](http://lirics.loria.fr/doc_pub/SynAF_WD_2006-01-22.pdf)

## 4.2 Term Finder (Keywords, Statistical and Linguistic)

Similar to Terminology Lookup, a term annotated as a 'DiscoveredTerm' can have annotations within them. For example, in the Figure 4 the potential term 'syntactic annotation' also has a potential term, 'annotation', within it. Discovered terms can also have known terms annotated within them. This new proposed term could then become an extension of the existing terminology. For example, in the figure above the discovered term 'dependency information' has the existing term 'information' annotated within it. Similarly, the proposed new term 'edge label' incorporates the known term 'label'. Decisions over the use of such relationships need to be considered. An example of how the terms were annotated within GATE for 'SynAF' is shown in figure 5.

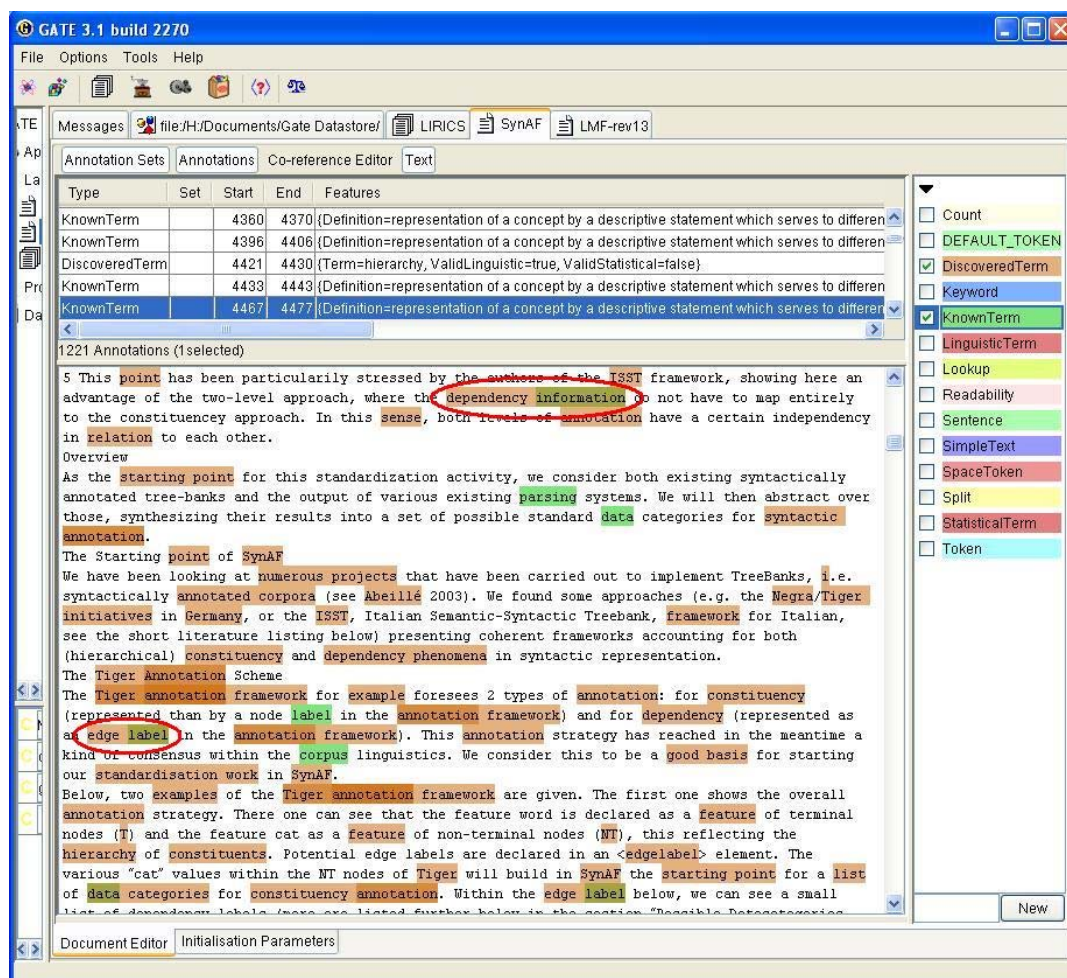


Figure 5: A screenshot of the 'SynAF' document in GATE displaying the KnownTerm and DiscoveredTerm annotations.

The numbers of known and discovered terms (total count) found in the two documents are detailed in Table 3. The 'LMF' document was roughly three times the size of 'SynAF', but appears to have substantially more terminological content.

Document	Known Terms	Discovered Terms
Lexical markup framework (LMF)	466	3712
Syntactic Annotation Framework (SynAF)	96	1125

Table 3: Number of known and potential terms in the ISO standards currently contained in the terminology database

The top 20 known terms in 'LMF' are shown with their frequencies in Table 4.

<b>Term</b>	<b>Count</b>
Paradigm	58
Data	56
label	41
extension	31
Text	29
word form	29
Information	26
Note	26
Definition	22
data category	21
Pattern	19
Type	17
Object	15
context	8
data category selection	8
Code	6
data category registry	6
homonymy	6
equivalence	4
String	4

**Table 4: The top 20 known terms, and their frequencies, in ‘LMF’.**

The top 10 known terms founds in the document ‘SynAF’ are shown in Table 5.

<b>Term</b>	<b>Count</b>
Type	28
label	15
data	14
definition	9
object	9
information	6
Merge	2
Parsing	2
Read	2
Context	1

**Table 5: The top 20 known terms, and their frequencies, in ‘LMF’.**

The discovered terms were further investigated to evaluate which could be considered as new terms. The terms highlighted as potential terms by both methods could be prioritised for consideration. Terms such as ‘syntactic annotation’, ‘annotation’, ‘SynAF’ and ‘morph’ were identified as items that may need to be defined. Further filtering of this list is required, but frequency information can be helpful here also; variations by part of speech can lead to duplications, for example for ‘SynAF’. Examples of discovered terms from SynAF are shown in Table 6.

Term	Linguistically Valid	Statistically Valid	Count
* annotation	false	true	42
head	true	false	33
value name	true	false	22
partec	true	true	21
* synaf	true	true	19
value	true	false	18
edge label	true	false	14
syntactic annotation	true	true	13
mod	true	false	11
morph	true	true	11
* synaf	false	true	11
word	true	false	11
* annotation	true	true	10
constituency	true	false	10
relation	true	false	10
data categories	true	false	9
Dependency	true	false	9

**Table 6: Examples of highly frequent discovered terms in ‘SynAF’, including duplications due to different parts of speech (\*).**

The discovered terms demonstrated, at lower frequencies, linguistically valid multiword expressions that were surprisingly complex. Examples of such terms are shown in figure 10.

Term	Linguistically Valid	Statistically Valid	Count
complex knowledge organization system	true	false	4
Imf data category selection procedures	true	false	2
semantic predicate class section	true	false	2
dual use mrd metamodel	true	false	2
dual use mrd package	true	false	2

**Table 7: Examples of potential multiword terms that were discovered in ‘LMF’**

Additionally, the linguistic and statistical methods for discovering terms found numerous valid two word expressions that were regularly used. Examples of these are shown in Table 8.

Term	Linguistically Valid	Statistically Valid	Count
process operator	True	false	11
affix allomorph	False	true	10
instance diagram	True	false	10
transform set	True	false	10
lemma writtenform	False	true	9
lexicon instance	True	false	8
syntactic argument	True	false	8
syntactic behaviour	True	false	8
xml fragment	True	false	8
external system	True	false	7

**Table 8: Examples of frequent discovered two-word terms in ‘LMF’.**

There were also notable keywords (single words) identified as valid either linguistically or statistically and frequently used throughout the document. Examples of these are shown in Table 9. Some of these may easily be filtered out.

Term	Linguistically Valid	Statistically Valid	Count
iso	true	true	77
lmf	false	true	41
subcategorization	false	true	37
synset	false	true	26
noun	true	false	24
sense	true	false	21
verb	true	false	20
word	true	false	19
english	true	false	18
lexicon	true	false	18

**Table 9: Examples of discovered single-word terms in ‘LMF’.**

The two methods of identifying potential new terms allowed for potential readability issues to be highlighted. Such a readability issue can be demonstrated by the first item in Table 7, the “complex knowledge organization system” – e.g. is it an “organization system” for “complex knowledge”, or a “system” for “complex knowledge organization” or is the “knowledge organization system” itself “complex”? Potential ambiguity could be demonstrated by specific differences in term recognition and entailment. The two terms shown in Table 10 were identified slightly differently by the two methods, and such differences may or may not be significant.

Term	Linguistically Valid	Statistically Valid	Count
multi-layered annotation	false	true	3
multi-layered annotation strategy	true	false	2

**Table 10: Complexity in entailed terms.**

The fact that the methods recognised the term differently demonstrated that it is not clear what is meant by a ‘multi-layered annotation strategy’. Is it a ‘strategy’ for a ‘multi-layered annotation’ or is it an ‘annotation strategy’ which is ‘multi-layered’. The two methods for recognising terms result in further considerations being required.

### 4.3 SimpleText Analyser

The first 150 suggested replacements from an earlier version of the LMF document<sup>3</sup> were manually analysed with 60 replacements being deemed appropriate. The result of this analysis was sent to the author of LMF and LIRICS project manager, Gil Francopoulo, to review these suggested replacements. The author agreed that 7 of these replacements were clearly appropriate, and the remaining 53 required further considerations to be made. A large proportion of the suggested replacements involved existing terminology. In an attempt to reduce the number of such false positives, the SimpleText analyser was, prior to the construction of the Annotation Controller, amended so that substitutions would no longer be suggested for text already annotated as a known term.

With the revised SimpleText analyser, a standard at comment stage, ISO/DIS 12620, was analysed to determine the potential for improvements. A report was produced of new replacements suggested by the SimpleText plug-in for words and phrases deemed unnecessarily complex. The first 200 replacements were analysed manually with the conclusion that there were 33 replacements that were suitable and of which there were 24 substitutions that were unique. Every further instance of the unique substitutions was analysed throughout the rest of the document, 183 instances in total, to see if the replacements were appropriate in every instance. These replacements and their results are detailed in Table 11.

---

<sup>3</sup> Revision 9, available at: [http://lirics.loria.fr/doc\\_pub/LMF%20rev9%2015March2006.pdf](http://lirics.loria.fr/doc_pub/LMF%20rev9%2015March2006.pdf)

Phrase	Replacement	Appearances In Document	Correct Replacements	%Correct
application	use	17	1	5.88%
by means of	by	2	2	100.00%
component	part	68	1	1.47%
comprises	is made up of	4	4	100.00%
consequence	result	1	1	100.00%
essential	important	2	2	100.00%
frequently	often	1	1	100.00%
in conjunction with	with	2	2	100.00%
in order to	to	4	4	100.00%
instances	cases	3	3	100.00%
latest	last	2	2	100.00%
nature	type	1	1	100.00%
needed	necessary	1	1	100.00%
permissible	allowed	4	4	100.00%
provide	give	19	3	15.79%
represent	show	6	2	33.33%
requirements	rules	4	2	50.00%
restrict	limit	1	1	100.00%
revised	changed	1	1	100.00%
specified	given	5	4	80.00%
thus	therefore	4	4	100.00%
utilize	use	1	1	100.00%
various	different	10	4	40.00%
within	in	20	14	70.00%

**Table 11: The 24 unique replacements filtered from the initial 200 suggestions with the number of times the replacements were correct throughout the rest of the document.**

#### 4.4 Readability analysis

It was found that some SimpleText replacements were appropriate in every further instance such as 'comprises', 'in order to', 'permissible' and 'thus'. However, other words rarely had correct replacements and in particular 'application' and 'component' were never suitable again. The majority of proposed SimpleText replacements were found not to be suitable and leaving much room for investigation in how to focus the substitutions more accurately. However, to investigate the extent that these limited number of substitutions would influence the readability scores of the document the Replacer plug-in was run. All the readability scores were slightly reduced except for the FOG and SMOG results which increased a little. The increase in readability scores can be attributed to the fact that some SimpleText replacements do not increase readability scores. In fact the number of words in a document can actually *increase* due to some of the replacements. The most common example of this occurrence is the substitution of 'comprises' for 'is made up of'. Other replacements such as 'important' for 'essential' has no effect on readability scores whatsoever as the number of syllables and characters is identical. The readability scores before and after the replacements are shown in figure 15.

Score	Before	After
Kincaid	18.411	18.367
Flesch	20.612	20.994
FOG	20.215	20.694
SMOG	16.993	17.436
ARI	18.027	17.947

**Table 12: Readability scores before and after the SimpleText process.**

Readability formulas have proven to be objective, easy to apply and quickly executed, making them useful for getting comparative measures of readability. However despite their advantages, they do not measure external factors that may make a text more easily understood, such as having access to a terminology, or the knowledge of words by general audiences. Definitions of readability state that a text should be compelling and comprehensible and the current readability measurements incorporate little of these elements into their calculations. As the tests currently stand, they provide little insight in how to improve the readability of text. Without further analysis of these aspects, these tests cannot be considered as definitive measures of readability.

## 4.5 Results from all Annotations

Information accumulated in the annotations by the plug-ins through the pipeline is shown in Figure 16. Annotations included in the figure are 'Count', 'Readability', 'KnownTerm', 'DiscoveredTerm' and 'Keyword'. The amount of information now being produced within these annotations is substantial, and further efforts are required to simplify the sets produced and to produce more human-readable assessments of the documents.

Type	Count	Features
Count	0 ...	{Characters=22129, PolysyllabicWords=838, Sentences=125, Syllables=7759, Words=4488}
Readability	0 ...	{ARI=19.74561631016043, Execution=2, FOG=21.830405704099825, Flesch=24.133215401069535, Kir
Readability	0 ...	{ARI=19.74561631016043, Execution=1, FOG=21.973008199643495, Flesch=24.133215401069535, Kir
DiscoveredTerm	...	{Term=annotation, ValidLinguistic=false, ValidStatistical=true}
Keyword	...	{FrequencyZ-Score=4.253701739544124, Text=annotation, WeirdnessZ-Score=0.49301052751493385}
KnownTerm	...	{Definition=information processing on language, ID=ID728, Substitute=No, Term=language processing}
SimpleText	...	{BestReplacement=-, Replacement1=show, Text=represent}
DiscoveredTerm	...	{Term=np, ValidLinguistic=true, ValidStatistical=false}
DiscoveredTerm	...	{Term=morpho-syntactically annotated items, ValidLinguistic=true, ValidStatistical=false}
DiscoveredTerm	...	{Term=morpho, ValidLinguistic=false, ValidStatistical=true}
Keyword	...	{FrequencyZ-Score=0.24944680300207364, Text=morpho, WeirdnessZ-Score=0.41329925163299047}
DiscoveredTerm	...	{Term=constituents, ValidLinguistic=true, ValidStatistical=false}
SimpleText	...	{BestReplacement=-, Replacement1=show, Text=represent}
DiscoveredTerm	...	{Term=dependency information, ValidLinguistic=true, ValidStatistical=false}
KnownTerm	...	{Definition=knowledge concerning such things as facts, concepts, objects, events, ideas, ID=ID324, Substi
SimpleText	...	{BestReplacement=-, Replacement1=is, Replacement2=are, Text=exist}
DiscoveredTerm	...	{Term=morpho, ValidLinguistic=false, ValidStatistical=true}
Keyword	...	{FrequencyZ-Score=0.24944680300207364, Text=morpho, WeirdnessZ-Score=0.41329925163299047}
DiscoveredTerm	...	{Term=morpho-syntactically annotated items, ValidLinguistic=true, ValidStatistical=false}
SimpleText	...	{BestReplacement=-, Replacement1=in, Replacement2=in less than, Text=within}

**Figure 6: Examples of annotations created by the document content management system.**

## 5 Discussion and further work

To provide for a variety of aspects of additional quality control, in addition to the extant processes of ISO, we have integrated and used a variety of supporting resources and components for the standards development process, including a Plain English thesaurus, lookup of ISO TC 37 terminology provided from a terminology management system (TMS) via ISO 16642, automatic terminology discovery using statistical and linguistic techniques, and readability metrics. These components have been re-engineered from the University of Surrey Department of Computing's content analysis applications (System Quirk), developed in prior research, including EU co-funded projects, and integrated with the University of Sheffield's GATE system. These efforts were undertaken to demonstrate the potential for controlled authoring in the International Standards environment. The result of these efforts leads us to the development of an assistive tool for authors of standards based around, and evaluated against, LIRICS work.

Initial experiments helped us to provide some additional commentary into ISO on a few standards documents at various stages of the ISO process; fuller sets of commentary for the LIRICS standards are at various stages of production and this deliverable presents some examples of how these can be formulated. Human interpretation of, and action upon, the

results being produced by these components is still required to varying extents, however the analysis of language simplicity and consistency, identification of known and unknown terms, and the generation of “understandability” metrics have all been implemented and demonstrate interesting and potentially highly-valuable results. Further evaluation efforts are needed to assess the results being produced, to improve the treatment provided and to improve the formulation of feedback on the document or documents being analysed. The ideal outputs would be fed directly to standards authors prior to the submission of a document into the ISO processes, potentially leading to a reduction in the quantity of comments relating to document syntax or terminology. Further work with standards authors to begin to embed the evaluation of the results into the authoring process is still required, and needed in LIRICS. Greater consideration of the management of overlaps between annotations is needed beyond this deliverable, and likely beyond LIRICS, and a number of future functional opportunities have been identified. Amongst these is a readability measure that takes full account of the processing, analysis and annotation sets outlined here.

The LIRICS proposal considered that, longer term, a content management system for developing standards was envisaged, but that this was considered beyond the scope of LIRICS. These efforts demonstrate a major step towards the provision of such a system and dealing with issues such as overlaps and inter-dependencies amongst the various components, and of providing further improved analysis, will require further efforts.

## References

Cunningham, H., Maynard, D., Bontcheva, K. and Tablan, V. (2002). "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications" in *Proceedings of the 40<sup>th</sup> Anniversary Meeting of the Association for Computational Linguistics(ACL'02)*. Philadelphia, July 2002.

English, H.B., and English, A.C. (1958) "A Comprehensive Dictionary of Psychological and Psychoanalytical Terms". London: Longmans, Green.

Flesch, R. (1948). "A new readability yardstick." *Journal of Applied Psychology* 32:221-223.

Gillam, L., Tariq, M. and Ahmad, K. (2007). "Terminology and the construction of ontology". In *Application-Driven Terminology Engineering*, Ibekwe-SanJuan, Fidelia, Anne Condamines and M. Teresa Cabré Castellví (eds.), pp49–73.

Gillam, L. and Ahmad, K. (2005). "Pattern mining across domain-specific text collections". *LNAI 3587*, pp 570-579

Gillam, L. (2004). "Systems of concepts and their extraction from text". Unpublished PhD thesis, University of Surrey

Gunning, R. (1952). "The Technique of Clear Writing". New York: McGraw-Hill.

Hargis, G., Hernandez, A., K., Hughes, P., Ramaker, J., Rouiller, S. and Wilde, E. (1998). "Developing quality technical information: A handbook for writers and editors". Upper Saddle River, NJ: Prentice Hall.

Jacquemin, C. (2001) "Spotting and Discovering Terms through Natural Language Processing". The MIT Press. ISBN 0-262-10085-1.

Kitson, H. D. (1921). "The mind of the buyer". New York: Macmillan.

Kincaid, J.P., Fishburne, R.P, Rogers, R.L. & Chissom, B.S. (1975). "Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for navy enlisted personnel". Research Branch Report 8-75. Naval Air Station, Memphis, TN.

McLaughlin, H. (1969). "SMOG grading - a new readability formula", *Journal of Reading*, 22, 639-646.

Smith, E. A. and R. J. Senter (1967). "Automated readability index", *AMRL-TR*, 66-22. Wright-Patterson AFB, OH: Aerospace Medical Division.