IPTC EXTRA Technical Requirements

Version 1.0 - 2017/01/30

1. Overview

- Participants: IPTC.org
- Status: Requirements Gathering.
- Target release: 2017

2. Goals and Objectives

Deliver an open source, rules based classification engine.

"Classification" means assigning one or more categories to the text of a news document. Rules based classifiers use a set of Boolean rules, rather than machine-learning or statistical techniques, to determine which categories to apply.

The rules consist of a Boolean expression and an associated term. If the Boolean expression evaluates to TRUE for a given document, then the associated term is assigned to the document. The Boolean expressions may consist of

- Boolean-valued operators
- Boolean-valued functions
- strings - which evaluate to “TRUE” if the string is present in the document

3. Background

For news publishers who are dissatisfied with either hand-tagging documents or statistical approaches to automated tagging, EXTRA is an open source, rules-based, classification system for annotating news documents with high-quality subject tags, regardless of language. Such tags allow publishers to deliver a variety of valuable services including content recommendations, improved advertising targeting and subject-specific content streams, such as alerts and topic pages.

Unlike hand-tagging, EXTRA’s rules-based system will allow publishers to tag their news content with consistent tags, at speed and at scale. Unlike statistical approaches, which often require numerous annotated examples, EXTRA’s rules-based system allows publishers to rapidly adapt to breaking news and low-frequency topics. Rules-based tagging is much more transparent than either manual tagging or statistical approaches. It is easier to explain - and alter - why a document is tagged for a particular topic in a rules-based system, whereas pure machine learning approaches are notoriously opaque.

To facilitate adoption and consistency, the IPTC will also create EXTRA extraction rules for tagging documents in two different languages with its industry standard Media Topics vocabulary.

4. Assumptions

4.1 Open Source

By “open source”, we specifically mean software licensed under the MIT License https://opensource.org/licenses/MIT

http://choosealicense.com/licenses/mit/ describes the MIT License as “A short and simple permissive license with conditions only requiring preservation of copyright and license notices. Licensed works, modifications, and larger works may be distributed under different terms and without source code.”

Any existing software that is used as part of the EXTRA project would therefore need to be licensed using the MIT license (or something that can be easily demonstrated to be compatible).

4.2 Technologies

It should be straightforward to run the EXTRA engine on a variety of modern hardware and operating system combinations. It should be possible to run the engine either on-premise or on well-known “cloud” platforms including Amazon Web Services, Google Cloud Platform and Microsoft Azure.

Candidate programming languages include Python, Java and C/C++ (each of which are widely used for Natural Language Processing applications and libraries). However, there is no strict requirement for the use of a particular language, as long as the other requirements and constraints are met. (On the other hand, use of one or more “mainstream” languages is preferred, since this implies easier support and maintenance in the future, all other things being equal).

4.3 Performance

4.3.1 Likely Performance Context

It is difficult to quantify performance requirements - and yet important. In an effort to provide some guidance, this section lays out likely “average” requirements.

It is likely that most news providers will have thousands of rules to evaluate per document. The majority will be relatively simple rules - such as name matches with a certain amount of disambiguating evidence - but some (perhaps about 20% of the total) will be quite complex, with multiple clauses and using several operators.

The typical document to be classified is likely to be 500-800 words long, although some will be much shorter and a minority will be much longer, in the typical scenario.

For the purposes of performance discussion, assume general-purpose hardware. At the time of writing, this might be an AWS m4.large https://aws.amazon.com/ec2/instance-types/ or a Google Compute n1-standard-2 https://cloud.google.com/compute/docs/machine-types.

4.3.1 Classify Content Response Time

Given ten thousand rules to evaluate, 95% of the time, EXTRA should respond with classification results for an average length document in a second or less on average hardware, under a sustained load of one classification request every two seconds.

4.3.2 Classify Content Scalability

It should be possible to scale the EXTRA classification capacity to handle more transactions in a given time period by adding additional hardware, without significant loss of performance. In other words, EXTRA should scale out - increase capacity by adding additional machines. (Increased performance by scaling up - making the machines more powerful - is fine too, as long as it is not the only option).

5. User Stories

The core workflows for using EXTRA are

- Creation, testing and maintenance of rules
- On-demand classification of content, such as when content gets created or published
- Bulk classification of content - such as from an archive

Secondary workflows, which support the development of rules, are

- Maintenance of document schema - the fields in the documents to be classified. The schema allow EXTRA to check that the field names in rules are valid
- Maintenance of dictionaries. The dictionaries allow EXTRA to check whether the terms used in the rules are spelt correctly.
- Maintenance of relevance algorithms. The relevance algorithms allow EXTRA to calculate a score for how well a given document matches a given rule.
- Highlighting “hits” and “misses”. Evaluate how well or poorly a given document matches a particular rule.

5.1 Create rule

- Create blank rule
- Duplicate existing rule
- Create rule from template
- Set metadata for the rule - at least the properties in Dublin Core must be supported http://dublincore.org/documents/dcmi-terms/

5.2 Update rule

- View rules
- Select rule
- Edit rule
- Edit metadata

5.3 Delete rule

- View rules
- Select rule
- Delete rule

5.4 Classify content

- Submit one or more documents to EXTRA and get back information about the rules which match each document.
- Submit one or more documents to EXTRA, specifying a root rule by id and for each document, get back information about which rules match - both the root rule and any helper rules. A helper rule is any rule which is referenced or triggered in order to satisfy the root rule.

5.5 Add schema

- Submit a new schema
- EXTRA returns an identifier for the new schema

5.6 Update schema

- Replace existing schema with a new schema

5.7 Delete schema

- Delete schema by id

5.8 Add dictionary

- Submit a new dictionary, with associated language metadata
- EXTRA returns an identifier for the new dictionary

5.9 Update dictionary

- Replace existing dictionary with a new dictionary

5.10 Delete dictionary

- Delete dictionary by id

5.11 Add relevance algorithm

See section 7.11 for a detailed discussion of “relevance”.

- Submit a new relevance algorithm, optionally with a range of rule ids. If no rule ids are specified, apply the algorithm to all rules, i.e. this is the new default.
- EXTRA returns an identifier for the new algorithm
- When both a default algorithm and a rule-specific algorithm have been defined, the rule-specific algorithm takes precedence
- It is up to the client to ensure that only one relevance algorithm is defined for a given specific rule id. If more than one is defined, then it is implementation dependent as to which one gets applied.

5.12 Update relevance algorithm

- Replace existing relevance algorithm with a new relevance algorithm

5.13 Delete relevance algorithm

- Delete relevance algorithm by id
- It is an error to attempt to delete the default relevance algorithm

5.14 Highlight hits

See section 7.3 for detailed discussion of hit highlighting.

- Submit a document to obtain “hits” for a given rule
- Returns the document with keyword hits marked up

5.15 Classification rule use cases

When writing rules using EXTRA, these are typical scenarios that need to be supported. Note that the EXTRA requirements should allow for each of these use cases. However, we felt it was useful to specifically list them here, for clarity.

- **Vary rules by document length** - see also 6.4.1.5 (rule parameters) and 7.4 Guideline #1 (relevance normalizing for document length).
- **Vary rules by document type** - such as handling documents which are composed of several different unrelated stories or handling photo captions or video scripts.
- **Disambiguation clauses**
- **Sentence pattern matching (grammar rules)**
- **Position in Document**: Make rules sensitive to position, higher weighting to occurrences in the top 10% vs. bottom 10%.
- **Recognition of proper names comprised of common noun parts**, e.g. don’t let Art Sherman trigger the rule for “Art”. Add negation to MINOC_5:”Art” that says not when followed by a title case word that repeats later in the article following Dr/Gen/Hon/Justice/Miss/Mr/Mrs/Ms/Rep/Sen/Sgt. (Note that these types of lists should ideally be specified once, so that they can reused in multiple rules, but maintained in one place).
- **Synonyms and gazetteers**, must be able to specify lists of name variants or vocabulary lists which are expressed as rules and so may be referenced by other rules.
- **Must allow global configuration of all rules.** Driven by variables - e.g. TEST or word count or … other things.
- **Aliases or macros.** A rule may reference another rule, for example to allow reuse of a rule in several rules. Such as to identify boilerplate text in a press release.
- **Rules can be collected together so that they can be triggered as a set.** For example, all rules to do with sports could be referenced by a single overarching sport rule, which you would use if you knew the content was only about sports.
- **Precision and recall testing.** Rules can be tested against a “gold” set of classified documents, to test the accuracy of the rules. This is typically measured using precision (when a document is classified by a particular rule, is it a correct classification) and recall (are all documents which should have been classified for a given term?) See also https://en.wikipedia.org/wiki/Precision_and_recall

6. User interaction and design

There should be a user interface for each of the functions of the EXTRA API, although, the UI is a lower priority than the API itself. A rich client UI, with the types of features outlined in section 6.5, is likely out of scope for the first phase of the EXTRA project.

6.1 All EXTRA features and defaults need to be documented

- All of the features and functions for rule writers need to be documented, with examples.

6.2 EXTRA Rules Language formats

- The EXTRA rules language format must be able to be expressed in either XML or JSON.
- Each rule may have a version number. If present, it will be returned when the rule matches.

6.3 Natural Languages

6.3.1 Supported Languages

The initial version of EXTRA must allow classification in (at least) these IPTC Media Topics languages:

- English
- French
- German
- Spanish

Amongst other things, this means that EXTRA must be able to differentiate the parts-of-speech for each language - including parsing into paragraphs, sentences and words. And identifying different word types, such as noun vs verb.

The initial version of EXTRA will support classification in at least two of the languages and will come with example rules in those languages, to demonstrate that support.

6.3.2 Character Encoding

Support for ISO/IEC 8859-1 (aka Latin-1) would likely be sufficient for supporting the Western European languages identified above. Ideally, however, EXTRA will support classification of a wide variety of languages. Therefore, EXTRA should support rules and documents encoded in UTF-8 Character Encoding, since that enables full Unicode support.

6.3.3 Dictionaries and Spell-Checking

Users of EXTRA may optionally manage a dictionary for each language. If a dictionary is supplied, then spell-check feedback may be requested for rules.

When updating or creating a rule, and spell check is requested, any words in the rule which are not found in the dictionary will be flagged as warnings.

6.4 Input Requirements

A document to be classified using the EXTRA engine may take a variety of forms. It may be structured, such as XML or JSON, or it may be unstructured, such as plain text. If it is structured, it may contain fields, such as ‘headline’ or ‘byline,’ which can then be targeted by the rule language.

In addition, the input document may be accompanied by additional metadata that will control which rules are applied and how these rules will behave. For example, we may wish to classify a set of documents against a subset of rules that are registered with the EXTRA engine. We may also wish to tune our classification to shorter or longer text, requiring each request to be accompanied by a parameter that can be used to guide this tuning.

6.4.1 Input Formats

To accommodate the above we have defined an input format, expressed in JSON, which contains both the document and its accompanying parameters. The document it contains can be expressed in XML, plain text, or in JSON. Certain formats will have limitations regarding what output features are available.

6.4.1.1 Example: An XML Document

```
{
    "document": "<body>This is some text to be classified.</body>"
}
```

6.4.1.2 Example: A Text Document

```
{
    "document": "This is some text to be classified."
}
```

6.4.1.3 Example: A JSON Document

```
{
    "document": "{\"body\":\"This is some text to be classified.\"}"
}
```

6.4.1.4 A document with rule selection parameters

It should be possible to select a set of rules for classification. If no rules are provided, all rules registered with the EXTRA engine are applied.

```
{
    "matches": [{
        "ruleid": "1234"
    }],
    "document": "<body>This is some text to be classified.</body>"
}
```

6.4.1.5 A document with rule selection and rule modification parameters

The way the rules are executed may be controlled with additional parameters supplied prior to classification. For example, we may wish to adjust the value of all occurrence operators based on expected document length. While this requirement shares a similar goal with 7.4 > Guideline #1, in that it alters the results with regard to document length, it differs in two respects:

- Parameter modification would occur pre-classification and modify the rule, based on an *assumption* of document length, but it would not actually calculate document length. 7.4 > Guideline #1, by contrast, would *not* modify the rule. It would calculate document length and adjust the relevancy score post-classification accordingly.
- Parameter modification would determine whether or not the rule matches. Whereas 7.4 > Guideline #1 would affect the relevancy score (assuming the rule does match) indicating the degree to which it matches.

So, for example, if the EXTRA engine were used to run classification on a collection of captions, which are assumed to contain few words (although precise document length is unknown), then the input parameters on all occurrence operators may be adjusted to trigger on a quarter (.25) of the normal threshold by writing.

```
{
    "parameters": {"minimum_occurrence": ".25"},
}
```

For rules requiring minimum occurrences of 4 and 8 respectively, as with the rules for American Football (rule id = 123) and Soccer (rule id = 456) below, the new parameter modification would reduce the threshold to 1 [.25 x 4] for American Football, and 2 [.25 x 8] for Soccer.

Rule id: 123
Rule name: American Football
Rule: (MINIMUM OCCURRENCE_4,"football","nfl","n.f.l.","super bowl")

Rule id: 456
Rule name: Soccer
Rule: (MINIMUM OCCURRENCE_8,"champions league","premier league","soccer","uefa")

Thus, these two inputs:

```
{
    "matches": [
        {"ruleid": "123"},
        {"ruleid": "456"}
    ],
    "parameters": {"minimum_occurrence": ".25"},
    "document": "<caption>NFL Roundup.</caption>"
}
```

```
{
    "matches": [
        {"ruleid": "123"},
        {"ruleid": "456"}
    ],
    "parameters": {"minimum_occurrence": ".25"},
    "document": "<caption>Real Madrid wins Champions
League.</caption>"
}
```

Would return these two outputs:

```
{
"meta":{"date":"2016-09-01T08:15:30-05:00Z"},
"matches":[{"ruleid":"123"}],
"document":"<caption> NFL Roundup.</caption>"
}
{
"meta":{"date":"2016-09-01T08:15:30-05:00Z"},
"matches":[{"ruleid":"null"}],
"document":"<caption>Real Madrid wins Champions League.</caption>"
}
```

The first caption would match rule 123, based on 1 occurrence of "nfl."
However, the second caption would fail both rules, with no occurrences of the text strings defined for rule 123, and only 1 occurrence of of the text string "champions league" for rule 456.

Since 1 occurrence is less than the required 2 [.25 x 8] required for a match of rule 456, no match is returned

6.4.1.6 A document with output controls

It should be possible to provide an array of output controls to request particular metadata in the output response.

6.4.1.6.1 Defaults

Defaults are `meta` (administrative metadata such as date of submission), `matches` (rule identifiers that matched the document), and `relevance` (how well a rule matches a given piece of content).

```
{
    "document": "<body>This is some text to be classified.</body>",
    "outputcontrols": ["meta","matches", "relevance"]
}
```

6.4.1.6.2 Returning input

Requesting `input` ensures that the entirety of the input data is returned as part of the output response.

```
{
    "document": "<body>This is some text to be classified.</body>",
    "outputcontrols": ["input"]
}
```

Particular `input` objects may also be requested:

```
{
    "document": "<body>This is some text to be classified.</body>",
    "outputcontrols":
["input.document","input.matches","input.parameters"]
}
```

If the input JSON contains a “meta” object, those values in the meta object will be returned in the output JSON as is. (Useful for tracing, comments and so on).

```
{
    "document": "<body>This is some text to be classified.<\/body>",
    "outputcontrols": [
```

```
        "input.document",
        "meta"
    ],
    "meta": {"echo.guid" : "239a8081d1af2ee0836fa8ac0fff1745",
"echo.creator" : "dfox"}
}
```

6.5 Output Requirements

The EXTRA engine accepts a document and returns classification metadata. At minimum this metadata contains the list of rules that were matched, but it may also contain other information, such as relevance scores per rule.

It is also possible to submit a document to EXTRA and request that only specific rules be considered for application. In this case, the output is the same as the above: a list of matched rules is returned (with optional relevance and other information).

Relevance indicates how well a rule matches a given piece of content and is expressed on a scale from 0 (irrelevant) to 1 (100% relevant).

Hit highlighting is limited to plain text and XML document bodies and is expressed using milestones - marking boundaries with empty elements, as described below.

To accommodate the above we have defined an output format which contains an array of rule matches, relevance scores, and the classified document with hit highlighting. The initial version is JSON.

6.5.1 A document

```
{
"meta":{"date":"2016-09-01T08:15:30-05:00Z"},
"matches":[{"ruleid":"1234"}],
"document":"<body>text</body>"
}
```

6.5.2 A document with relevance

```
{
"meta":{"date":"2016-09-01T08:15:30-05:00Z"},
"matches":[{"ruleid":"1234","relevance":1.0}],
"document":"<body>text</body>"
}
```

6.7 Richer UI Suggestions

These are some ideas for the features that a richer User Interface for EXTRA should support

- View rules
    - Collapsable/expandable tree directories
    - Browse by category and date created/modified
    - Search by keyword
    - Visually represent rule hierarchy
- Select rule by using sorting and filtering options
- Batch update rules of the same type/template
- Undo editing
- Batch delete
- Undo delete
- Drag-n-drop

6.8 Error, warning and informational messages

The API will support HTTP error codes. Additional information for particular error conditions should be supplied whenever possible.

When creating or updating a rule, operator names which are incorrect will be flagged as errors. Syntax errors will be flagged as errors. Field names which are not known to EXTRA will be flagged as warnings.

7. Rule Language Requirements

7.1 Boolean Operators and Functions

The Boolean expressions may use the following operators and functions.

Operator - a short name for the operator or function
Definition - a brief description describing what the operator does and how it should be used
Priority - how important is it that the operator is supported by EXTRA
XQuery FT - the equivalent operator (if any) in XQuery Full Text https://www.w3.org/TR/xpath-full-text-10/
ElasticSearch - the equivalent operator (if any) in ElasticSearch DSL
https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html
Teragram - the equivalent operator (if any) in SAS Teragram https://www.sas.com/en_us/software/teragram.html

Operator	Definition	Priority	XQuery FT	ElasticSearch	Teragram
AND	Takes two or more statements. Category	MUST	ftand	{"bool":{"must"}}	AND

	matches only if all the statements are true.				
OR	Takes two or more statements. Category matches if at least one statement is true.	MUST	ftor	{"bool":{"should"}}	OR
NOT	Used in combination with AND. Category matches if the statement does not appear in combination with statement under the AND.	MUST	ftnot	{"bool":{"must_not"}}	NOT
MINIMUM	Combined with a number (e.g., MINIMUM_2) Takes one or more statements. Category matches if a minimum of x statements from the list appear in the text.	MUST			MIN_
DISTANCE	Combined with a number. Takes two or more statements. Category matches if statements are within x number of words from each other.	MUST	distance (exactly|at least|at most|from X to Y)	span_near with slop param	DIST
MINIMUM OCCURRENCE	Combined with a number. Takes one or more statements. Category matches if statement appears x amount of times in the text.	MUST	{} any word occurs at least N times	minimum_should_match (parameter on a bool query)	MINOC_
ORDER	Takes two or more statements. Category matches if statements appear in the text in the same order that they appear in the rule.	MUST	ordered	span query with in_order param set to "true"	ORD
SENTENCE	Takes two or more statements. Category matches if statements appear within the same sentence.	MUST	same sentence	nested AND queries (sentences need to be indexed as sub-fields in ES)	SENT
NOT WITHIN DISTANCE	Combined with a number. Takes two or more statements.	MUST	($a ftand $b) ftand ftnot ($a	span_not with dist param set	NOTINDIST_

	Category matches if the statements are not within x amount of words from each other.		ftand $b distance exactly N words)		
PARAGRAPH	Takes two or more statements. Category matches if all the statements occur in the same paragraph.	MUST	same paragraph	nested AND queries (paragraphs need to be indexed as sub-fields in ES)	PAR
NOT IN PHRASE	Takes two statements. Category matches if the first statement occurs outside of the second statement.	MUST	not in	span_not	NOTIN
NOT IN SENTENCE	Takes two or more statements. Category matches if all the statements do appear in the same document, but not in the same sentence.	MUST	different sentence	nested AND queries (sentences need to be indexed as sub-fields in ES)	NOTINSENT
NOT IN PARAGRAPH	Takes two or more statements. Category matches if all the statements do appear in the same document, but not in the same paragraph.	MUST	different paragraph	nested AND queries (paragraphs need to be indexed as sub-fields in ES)	NOTINPAR
ORDER AND DISTANCE	Combined with a number. Takes two or more statements. Category matches if both statements occur in the same order in which they are written in the rule and if both are within x amount of words to each other.	MUST	distance (exactly|at least|at most|from X to Y) ordered	span_near with in_order param set to "true"	ORDDIST_
MAXIMUM OCCURRENCE	Combined with a number. Takes at least one statement. Category matches if the statement appears no more than x amount of times in the text.	SHOULD	{} any word occurs at most N times	{ "query": { "bool": { "minimum_should_match": [MINIMUM], "should": [your queries], "must_not": ["bool": {	MAXOC_

				"minimum_should_match": [MAXIMUM – 1] "should": [your queries again] }] } } }	
FROM START	Combined with a number. Takes one or more statements. Category matches if the statement appears within x amount of words from the start of the text.	SHOULD			START_
FROM END	Combined with a number. Takes one or more statements. Category matches if the statement appears within x amount of words from the end of the text.	SHOULD			END_
MAXIMUM SENTENCES	Combined with a number. Takes one or more statements. Category matches if the statements appear with the first x sentences.	MAYBE			MAXSENT_
MAXIMUM PARAGRAPHS	Combined with a number. Takes one or more statements. Category matches if statement appears within the first x paragraphs.	MAYBE			MAXPAR_
PARAGRAPH POSITION	Combined with a number. Takes one or more statements. Category matches if statement appears within the first x paragraphs.	MAYBE			PARPOS_

7.2 Rule Language Requirements

7.2.1 Rules must be able to reference other rules

Rules may be Boolean combinations of other rules. Rules may be referenced are by rule ID.

7.2.2 EXTRA Must be able to support large numbers of rules
Over two million.

7.2.3 Rules MUST support parameters
The EXTRA engine must supply a set of parameters to the rules, the values of which can then be used by the rules. For example, word count and number of characters.

7.2.4 Rules MUST support IF THEN ELSE clauses

7.2.5 Rules MUST support Regular Expressions
Ideally, full PERL regular expressions would be supported https://en.wikipedia.org/wiki/Regular_expression#Perl. But at least POSIX Extended would be ideal https://en.wikipedia.org/wiki/Regular_expression#POSIX_basic_and_extended

7.2.6 Rules MUST support calculations
For example, to calculate 10% of a document length

7.2.7 Rules must be able to reference variables set in other rules
For example, to set a parameter based on document length.

7.2.8 All operators which take a number parameter MUST be able to be to take a variable as a value.
For example, the MINIMUM operator may take as a number value a variable which has been set in another rule. (This may make efficient evaluation of the rules hard - but we'd like to find out whether that is the case before we drop this as a requirement, since document length is key to the effectiveness of many classification rules).

7.2.9 Rules MUST support the ability to apply stemming to statements in supported languages
And optionally to choose - at minimum - verb or noun or all stems. Stemming should always follow part-of-speech recognition.

7.2.10 Rules MUST support nesting of operators
Enable nesting of operators. Example: DISTANCE operator inside MINIMUM_OCCURRENCE.

7.2.11 Rules MUST support scoring and the ability to weight the scores
Algorithm TBD but something like Score = How often the rule statements appear in the document divided by the number of terms in the document. Rule writer needs to be able to specify weighting per field. Either globally (all rules) or per rule.

7.2.12 Rules MUST support the ability to specify case sensitivity

7.2.13 Rules MUST support the ability to specify ascii folding
Rules may specify that the document to be classified must be ASCII folded meaning: convert alphabetic, numeric, and symbolic Unicode characters which are not in the first 127 ASCII characters (the "Basic Latin" Unicode block) into their ASCII equivalents, if one exists.
For example, to convert José -> Jose, Pâté -> Pate, Pokémon -> Pokemon. For example, using the conversions specified in
https://github.com/tnajdek/ASCII--Dammit/blob/master/AsciiDammit.py

7.3 Hit Highlighting

When writing and maintaining rules, it is useful to be able to see how well a particular document matches a particular rule - and vice versa. EXTRA supports this by offering a hit highlighting service. Note that, in order to support the hit highlighting described below, the EXTRA classification service must return indicators of which parts of a rule match a document and, by implication, which parts do not match. In particular, if a NOT clause is satisfied, this will be highlighted in the hit highlighting markup.

7.3.1 Marking up Hits Using Milestones

To markup the rule hits within a document, EXTRA uses TEI's <milestone> XML element:

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/CO.html#CORS5
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-milestone.html

Specifically, the start and end of a rule clause match is indicated by a pair of milestone elements. Each milestone element may use the following attributes:

- n - the clause which is being delimited by the milestone elements. No need to be unique. The value of @n indicates which part of the rule matches this part of the document. Required.
- xml:id - the unique identifier for the element. Required.
- spanTo - used by the start milestone to identify the end milestone by its xml:id value. Required on the start milestone.

```
<milestone n="Charlotte" xml:id="m1" spanTo="#m2"/>Charlotte<milestone n="Charlotte"
xml:id="m2"/>
```

It is quite likely that clause matches in a given document will overlap. The milestone mechanism supports these overlaps, including a string which matches more than one clause.

Please see Appendix B for examples of rules and documents marked up to indicate hits.

7.3.2 Marking up Hits Using Stand-off Markup

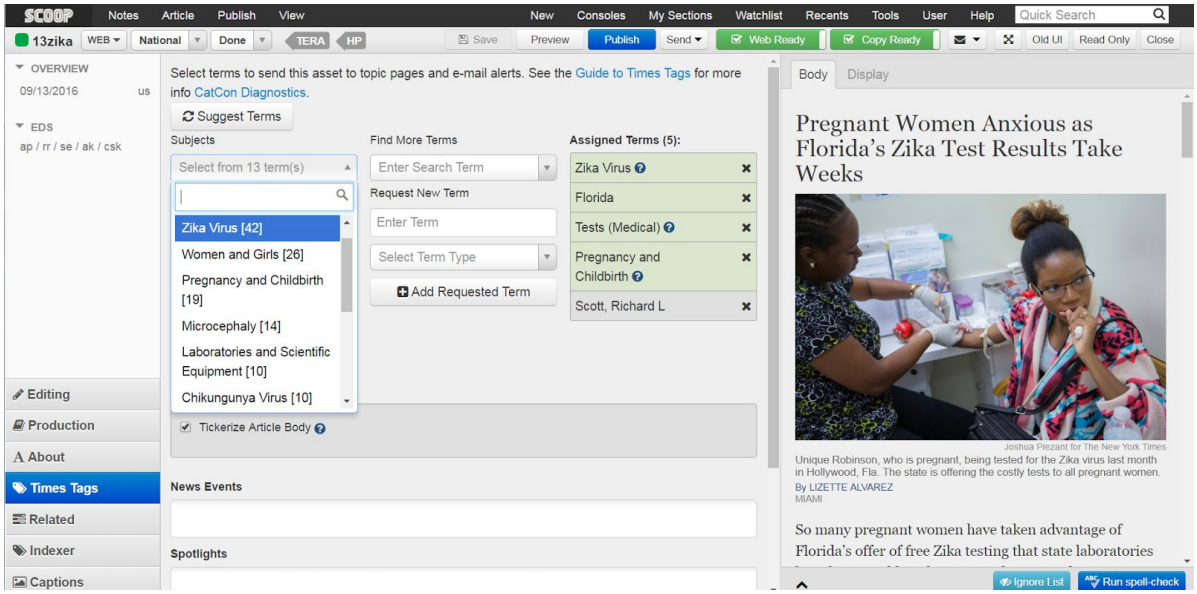
The alternative to altering the original document using milestones to indicate the location of hits is to use “stand-off” markup http://wiki.tei-c.org/index.php/Stand-off_markup

Stand-off markup is the opposite of inline markup: it can be used to indicate character ranges and associate metadata with those ranges. In this case, it could be used to link rules with documents. It is equally as expressive as the milestones mechanism and so either approach is acceptable.

7.4 Relevance

When classifying a document, EXTRA calculates and returns a score of how well each rule matches a given piece of content. While the rule determines ***whether or not*** the document matches a given topic: 0 (no) and 1 (yes), it is the relevance algorithm that indicates ***how strongly*** the document matches the topic: 0 (irrelevant) to 1 (100% relevant).

This score may serve as a valuable data point to rank rule results for human selection and/or to set a threshold for automated selection by software. For example, the following screenshot illustrates a CMS implementation that ranks terms by relevance for editorial selection. In an article about the Zika virus, in which Chikungunya is mentioned peripherally but is not the main focus, the term "Zika Virus" is ranked with a score of 42, whereas "Chikungunya Virus" receives a score of 10. Given these scores, Zika appears higher than Chikungunya in the pulldown for selection.



7.4 Relevance Guidelines. While the relevance algorithm should be a collaboration between the developer and rule writers, the guidelines below may be used as a starting point of discussion. The following list is intended to enumerate all possible requirements. Specifically, we would like to have a default algorithm, built into the EXTRA engine, with the ability for rule writers to override the default with a custom engine. (The factors which should be customizable are listed as “custom”, with all others listed as “default). We have also identified those relevance factors which we believe are most important to include within the engine (“core”), together with others that we will feel would be powerful but only “nice to have”. For an overall matrix of requirements belonging to each category, please refer to 7.4.4.

1. **Frequency with regard to Document Length, Type of Material and Occurrence Operators:** The EXTRA algorithm should use term frequency to boost relevance, while normalizing the results with regard to **document length**. So when a document matches a rule, the relevance should be boosted by the frequency of matches within the document. However, since documents vary in length -- with matches likely to appear more often in longer documents than shorter ones -- term frequency should be divided the document length in order to normalize results. Document length, for the purposes of this document, is defined by word count. The definition of a word -- whether orthographic, lexical, content or grammatical etc. -- is left to the developer to advise which approach will maximize relevant returns.

   Similar to document length, frequency may also be normalized for **type of material**, although on customizable basis (see Requirements Matrix in 7.4.4) since type of material may vary significantly across organizations. Sample types of material may include article, brief, anthology, photo, caption, slideshow, video, etc. While this requirement may seem similar to document length, there is a significant distinction. Document length measures straight word count. Type of material takes into account that the asset may not be not standalone and therefore, represents a smaller component of content belonging to a larger whole. As such, a match on a component (vs. composite) asset might be considered less likely to occur and when it does, it warrants a higher score. For example, a photo caption and a news briefing both contain 25 words. A rule match on the caption, which describes one photo in a longer story, should weigh more heavily than a match on the briefing, which -- while it may be short -- contains the full breadth of a beginning-middle-and-end and represents a self-contained content unit.

   Additionally, **occurrence operators** (such as MINIMUM OCCURRENCE, or MAXIMUM OCCURRENCE) should factor into the equation, by boosting the relevance by a factor proportional to how often the document matches the occurrence operator. For example, if rule A contains a MINIMUM OCCURRENCE=10 and rule B a MINIMUM OCCURRENCE=2, and the document matches both rule A & B with an occurrence of 20 for the required phrase in each, then rule B should score higher than A. That is, 20 occurrences matches rule A by a factor of 2, but it matches rule B by a factor of 10.

Note that while the frequency requirements above may draw from the "term frequency" portion of TF/IDF, the IDF property should not be applied. That is, IDF measures how frequently a word appears in many documents across a collection (with a high frequency indicating it is less unique and scaling down the importance of stop words like a/an/the/for). However, since EXTRA will be comparing the relevance of one rule vs another across a single document, not a collection of documents, document-to-document relevancy is not appropriate for this algorithm.

2. **Position in the asset:** The relevance algorithm should take into the account the position of where matches occur in the document. Positioning may be determined by either word count or by fields.
    a. **By word count:** Matches that occur closer to the beginning of the document should be assigned higher relevance than those that occur toward the end. Matches that occur at both the beginning and the end, should receive the highest score.
    b. **By fields:** For fielded input, EXTRA should be able to assign higher weighting to fields that have been designated by the rule writer as having greater importance. These assignments may be variable per rule since different rules key off of different fields.

        In order to illustrate some examples, it is helpful to first define a glossary of fields that may be used by news providers:

Body	The text beginning with the first word in the first paragraph to the last word in the final paragraph
Byline	The author of the asset
Dateline	The date and location where the reporting occurred
Kicker	A short phrase that precedes the headline and designates a collection of stories, such as an ongoing column or series
Headline	The title of the asset
Lede Graph	The first paragraph
Section	A label for site navigation that groups content topically
Subsection	A label for site navigation that sits hierarchically under section.
Summary	An abstract of 1-2 sentences that either summarizes the content of the asset, or extracts 1-2 key sentence(s) to entice the reader to read more

Type of Material	The structural template of the asset, e.g. news article, review, editorial, op-ed, slideshow, photo, video

Using these fields as structured input, a rule writer may designate certain fields as more important than others on a per-rule basis in order to optimize relevance. The following excerpts of rules for "Movies," "Obituaries," and "European Sovereign Debt Crisis" illustrate some examples.

Rule = Movies. Some reporters, such as a movies columnist, may write exclusively on a single topic. In this scenario, the rule for "Movies" should weight byline highest. Next might be "Section," which might alternate between "Movies" on weekdays and "Weekend Arts" on weekend, followed by "Summary","Kicker","Headline," and then "Body."

Sample Input:

Byline: Manohla Dargis
Section: Movies
Kicker: Review
Summary: “Sully,” Clint Eastwood’s film about US Airways Flight 1549, largely involves what happens after he, his crew and his passengers are plucked from the Hudson.
Headline: Sully Landed the Plane. Then He Had to Endure the Spotlight.

Rule part:

(OR,_byline:"manohla dargis",_section:"movies",_summary:"film@",_summary:"movie@",(OR,_headline:"academy award",_headline:"academy awards",_...,(OR,(OR,"rated pg-13","rated r","rated pg","rated g","under 17 requires accompanying parent")

Rule = Obituaries. For some news providers, obituaries may adhere to a certain structure. For example, in a New York Times obituary, the first sentence states that someone has died, and the second sentence states the age of the deceased. Given this structure, the lede should be weighted highest.

Sample Input:

Lede Graph: Robert E. Allen, who as chief executive and chairman of AT&T for almost a decade presided over major reorganizations of the company as the telecommunications industry remade itself in the late 1980s and ’90s, died on Saturday in Chatham, N.J. He was 81.
Headline: Robert E. Allen, 81, Dies; Led an AT&T in Transition

Summary: Mr. Allen was chief executive and chairman of the communications giant in the 1980s and ’90s as it acquired wireless and computer businesses.

Rule part:

(ORD,(SENT,(OR,_lede:"died",_lede:"found dead")),(SENT,(OR,_lede:"he was",_lede:"she was")))

Rule = European Sovereign Debt Crisis. A series on a particular topic may use a consistent kicker label to denote each installment. For example, an ongoing series on the debt crisis in Europe might use a kicker label "The Debt Crisis." Matching this field would be a direct hit for the corresponding rule on this topic. Thus, for this rule, the kicker should be weighted highest.

Sample Input:

Kicker: The Debt Crisis
Summary: Italy accounts for a third of the eurozone’s nonperforming loans. But that hasn’t stopped its banks from extending credit to loss-making companies.
Headline: Italian Banks Continue to Lend to Stagnant Companies as Debt Pile Mounts
Lede Graph: In Italy, where two decades of economic stagnation have created a long line of barely breathing companies, Feltrinelli, one of the country’s largest booksellers, stands out.
Subsection: Europe
Section: World

Rule Part

_kicker:"Debt Crisis", (SENT,(OR,"austerity measures","austerity plan","austerity program","bailout","debt crisis","financial crisis","financial rescue"

3. **Weighting of Operators:** Similar to weighting of fields, the weighting of operators may optimize results, especially for news providers who may not have structured input to leverage weighting by fields.

    a. Weighting per rule: Within a rule, a rule-writer may specify that certain operators take precedence over others. Returning to the Obituaries example in 2b, we could achieve similar results by weighting the operators ORDER and SENTENCE highest.
    b. Weighting across rules: A rule-writer may specify that operators that take two or more statements -- AND, DISTANCE, ORDER, PARAGRAPH, SENTENCE --

are weighted more strongly across all rules than those that take only one -- OR, MINIMUM, MINIMUM OCCURRENCE.

4. **Weighting of Text Strings:** Sometimes a match on a singular text string -- without regard for operator, or field, or placement in the text, simply the string itself -- might constitute a strong match. In these cases, it would be useful to assign a high weight to that string.

   For example a very specific court case, might be sufficient evidence to trigger a rule. The text string "Gideon v. Wainwright" -- a landmark cases requiring indigent criminal defendants be provided legal counsel -- would be enough to return a positive result on the rule "Public Defenders and Court-Appointed Lawyers (Criminal)."

5. **Rule Expirations:** For scheduled events that have a definitive end date, it would be useful to weight the rule according to an expiration date. For example, the rule for the 2016 Olympics may assign greater weighting to any matches that occur in the months leading up to the Opening Ceremonies, and then fall following the Closing Ceremonies. Or for the 2016 Presidential Election, "Clinton" or "Trump" may weigh more heavily in the lead-up to the election.

6. **Reduced weight to asynchronous or anomalous entities, or metonyms:** To help parse out the use of figurative language that may confuse the engine, it may be useful to assign lesser weightings, or flag for editorial review, to the entities below. Whether this goal may be achieved through actual entity recognition (and thus, customizable per language) or more computationally by recognizing outliers in otherwise homogenous content, is left to the discretion of the developer.
   a. Asynchronous or Anomalous entities.
      - Asynchronous: Historical references when they appear in the context of current events are often not the primary focus of the article. For example, articles that refer to Afghanistan as the next Vietnam, or Gary Johnson's third party campaign when compared to Ross Perot's. In these cases, the article is not about Vietnam or Ross Perot, but simply using them as points of reference. For this reason, they should receive low relevance.
      - Anomalous: Metaphorical references that standout in otherwise homogenous content, should be flagged as low relevance. For example, a Gail Collins column comparing Barack Obama's low poll numbers to the losing streak of the NY Mets baseball team. In this instance, the bulk of the content deals with US Politics, but there is a sprinkling of references to baseball. The baseball references should receive a lower weighting, or even cancel out altogether.
   b. Metonyms that don't carry a literal meaning when used in a news context can cause confusion. For example, news articles often refer to capital cities as shorthand for the seats of government. When Beijing appears in proximity to Xi

Jinping, the reference is most likely not to the city of Beijing, but to the Chinese government. Other common metonyms include: Broadway (for theater), Brussels (for the EU), Davos (for World Economic Forum), Detroit (for US auto industry), Downing Street (for British government), Hague (for various tribunals), Hollywood (for US movie industry), Silicon Valley (for US tech industry) and the Vatican (for the Roman Catholic church).

7.4.1 EXTRA will include a default relevance algorithm. The default algorithm should include the features marked in the matrix in 7.4.4.

7.4.2 The default relevance algorithm can be overridden with a different default algorithm. A user of EXTRA should have the option to override the default relevance algorithm with a custom one if necessary.

7.4.3 The default relevance algorithm can be overridden per rule. Different rules key off of different properties. As such, it may be necessary to use different relevance algorithms for different rules, so that relevance scores may be compared between topics.

7.4.4 Requirements Matrix. The following matrix organizes the requirements according to whether they are considered part of the default algorithm or customizable by rule, as well as whether they are considered core functionality for the initial prototype, or nice-to-have's for some future release.

Weighting Requirement	**Default**	**Custom**	**Core**	**Nice-to-Have**
Frequency by Word Count	x		x	
Frequency by Type of Material		x		x
Frequency by Occurrence Operator	x		x	
Position by Word Count	x		x	
Position by Field		x	x	
Operators per Rule		x	x	
Operators across Rules		x	x	
Text Strings		x		x
Rule Expirations		x		x
Asynchronous/Anomalous		x		x

Entities & Metonyms				

8. EXTRA and Machine Learning

EXTRA is a rules-based classification engine. The primary purpose of the engine is to allow people to create and maintain rules for classifying content. But is there any role for Machine Learning? Some possibilities include...

8.1 Getting a Head Start: Computer-Aided Rule Writing

It may be possible to exploit Machine Learning techniques to accelerate the rule-writing process, given access to a news corpus.

8.1.1 Generating a Candidate Set of Rules from an Annotated Corpus

If the rule writer has access to a news corpus which has already been annotated with the desired tags, then a ML module could derive an initial set of rules, which can then be perfected by the rule writer. The rules could be derived by, for example, identifying the ngrams uniquely associated with each topic. This is an approach similar to the one implemented in https://github.com/tmadl/sklearn-expertsys

8.1.2 Semi-Supervised Generation of Rules from an Unannotated Corpus

If the rule writer has access to a news corpus which doesn't already have the tags applied, it may still be possible to use Machine Learning techniques. For example, active learning techniques could be used to interactively work with the rule writer to determine the correct tags to apply - and hence be able to generate a candidate rule set - without the need to manually tag the entire news corpus.

8.2 Automatic Maintenance of Concepts and Terminology

Part of the art of rule writing is to identify concepts by the terminology associated with them. Sometimes, these concepts are used as evidence for a topic being relevant, but they may also be used to disambiguate topics, i.e. to identify concepts which might be confused and to ensure that the wrong terminology doesn't occur in the document.

Maintaining the up-to-date set of terminology associated with a given topic requires eternal vigilance. It is possible that ML techniques - such as statistical classification, named entity recognition and co-occurrence identification - could help with the maintenance of concepts. If the maintenance was applied automatically, then the rule writer would essentially have a

higher-level language to work with. Alternatively, the ML techniques could suggest edits to rules, by monitoring classified documents and identifying new candidate terminology.

9. Nice to Have

The following are items which have been discussed during requirements gathering. They have been judged as either less importance or potentially too complex to implement during an initial version of EXTRA. Naturally, if we can accommodate, that would be ideal. However, it could well be that there are additional rounds of development (and perhaps funding) where some or all of the following might be addressed.

- Document section relevance: EXTRA needs to support scoring of how well a rule matches by section. Optionally apply rules to section of documents, with a means to identify each section - e.g. paragraph marker. You tell the categorizer how to define a section, and then target rules based on that. This could mean that the rules vary based on length or other factors that are relevant for sections shorter than the entire document.
- Tracking, log of user activity
- Testing rules - run a set of scored content against the rule. Return precision and recall scores for the rule against the scored content set. Return list of documents which now match and didn't before. Return list of documents which no longer match and did before. It is recommended that during pilot testing or possibly as ongoing quality assurance, some subset of classified documents to be verified as accurate by human reader(s).
- Event extraction - extract details of events (e.g. <CompanyA> <acquires> <CompanyB>, <Company> <hires> <Executive>) - perhaps based on templates?
- Adapters allowing EXTRA to be easily integrated with popular CMSes and editorial tools.
- Arabic is an IPTC Media Topics language, but typically has poor NLP support, compared to Western European languages.
- Inline markup of entities, to associate a link with an entity. Or perhaps keep serialization separate? Would need to be optional.
- Identify other rules which are similar

10. Out of Scope

1. Versioning of individual rules by EXTRA is out of scope.
    a. Because there are other ways to version rules outside of EXTRA.
    b. Because it is way too complicated to support invoking particular versions of rules.
2. Filtering or searching using the metadata associated with the rules is out of scope.
3. Enforcement of entitlements - initial version just allows permitted users access to all features of the EXTRA engine.
4. Automated conversion of rules from another language into EXTRA's rule language
5. Targeting part of documents - e.g. to handle list of stories, chapters within books, news digests

Appendix A: Sample Rules

A.1 Police Brutality, Misconduct and Shootings

Contains examples of ORDIST with nested NOT's to differentiate active vs. passive (cops shot vs. cops were shot).

```
(OR,
   (AND,
      (OR,
         (OR,"beaten by police"," black deaths at the hands of white police
officers","charges against two troopers","convicted officer","convicted
officers","convicted trooper","convicted troopers","officer shoots","shot by
cop","shot by a cop","shot by cops","shot by police","shot by a police","shot by a
trooper","shot by troopers","shot by an undercover police","shot by an undercover New
York City","shot by undercover",
            (AND,"police",
               (OR,"disciplinary action","downgrading criminal
complaints","incorrectly classifying crimes","officer charged in","officer charged
with","officers charged in","officers charged with","officer convicted in","officer
convicted with","officers convicted in","officers convicted with","officer facing
charges","officers facing charges","officer was charged","officers were
charged","officer was convicted","officers were convicted","wrongdoing"
               )
            ),"police abuse","police behavior","police brutality","police
conduct","police corruption","police harassment","police had beaten","police had
overreacted","police misconduct","police officer shot","police officers had
violated","police officers of beating","police officers shot and wounded",
            (OR,"police shoot","police shooting","police shoots","police shot","police
shootings"
            ),"policemen beating","stop-and-frisk practices"
         ),
         (DIST_2,
            (OR,
               (AND,"officer",
                  (NOT,
                     (OR,"correction officer","corrections officer"
                     )
                  )
               ),
               (AND,"officers",
                  (NOT,
                     (OR,"correction officers","corrections officers"
                     )
                  )
               ),"police","policeman","policemen","policewoman","policewomen"
            ),
```

```
            (OR,"charged with","facing charges","suspended"
            )
         ),
         (SENT,
            (OR,
               (AND,
                  (OR,"officer"
                  ),
                  (NOT,
                     (OR,"correction officer","corrections officer"
                     )
                  )
               ),
               (AND,
                  (OR,"officers"
                  ),
                  (NOT,
                     (OR,"correction officers","corrections officers"
                     )
                  )
               ),"police","policeman","policemen","policewoman","policewomen"
            ),

(OR,"beat","beaten","brutalize","brutalized","brutality","chokehold","chokeholds","cho
ke hold","choke holds","choke-hold","choke-holds","corrupt","corruption","excessive
force","excessive use of force","fixing a ticket","fixing tickets","fixed a
ticket","fixed tickets","internal affairs investigation","misconduct","racist
email","racist emails","racist memes","racist
texts","sodomize","sodomized","sodomizing","ticket-fixing","ticket fixed","unnecessary
force","unnecessary use of force"
            )
         ),
         (AND,
            (OR,
               _headline:"police",
               (OR,"police"
               )
            ),
            (OR,
               (MINOC_2,"fixing a ticket","fix a ticket","fixed a ticket","fix
tickets","fixed tickets","fixing tickets","ticket-fixing","ticket fixed"
               ),
               (SENT,

(OR,"falsify","falsified","manipulate","manipulated","manipulating","manipulation"
                  ),
                  (OR,"crime report","crime reports","crime statistics"
                  )
               ),
               (AND,"eli b. silverman","john a. eterno"
```

```
               ),
               (OR,"mistake by an officer","mistake by officers","mistake by
trooper","mistake by troopers","officer faked","officers faked","trooper
faked","troopers faked","offices were fired","officer was fired","officer fired
for","officers fired for","troopers were fired","trooper was fired","trooper fired
for","troopers fired for"
               )
            )
         ),
         (AND,"louis scarcella",
            (OR,"detective","police","nypd","n.y.p.d."
            )
         ),
         (ORDDIST_3,
            (OR,"agent","agents",
               (AND,"officer",
                  (NOT,
                     (OR,"correction officer","corrections officer"
                     )
                  )
               ),
               (AND,"officers",
                  (NOT,
                     (OR,"correction officer","corrections officers"
                     )
                  )
               ),"cop","cops","police","policeman","policewoman","state
trooper","undercover New York City","undercover officers","undercover police"
            ),
            (OR,"beat",
               (AND,"fatally shot",
                  (NOT,
                     (OR,"was fatally shot","were fatally shot"
                     )
                  )
               ),
               (AND,"killed",
                  (NOT,
                     (OR,"was killed","were killed"
                     )
                  )
               ),"killing",
               (AND,"shot",
                  (NOT,
                     (OR,"was shot","were shot"
                     )
                  )
               ),"shooting"
            )
         ),
```

```
         (ORDDIST_3,
            (OR,"beat","beaten","fatally shot","killed","killing","shot","shooting"
            ),"by",
            (OR,"agent","agents",
               (AND,"officer",
                  (NOT,
                     (OR,"correction officer","corrections officer"
                     )
                  )
               ),
               (AND,"officers",
                  (NOT,
                     (OR,"correction officer","corrections officers"
                     )
                  )
               ),"cop","cops","police","policeman","policewoman","state
trooper","undercover"
            )
         ),
         (AND,"police",
            (OR,"abner louima","donnell mcfarland","eric garner","freddie gray","jamar
clark",
               (AND,"michael brown","ferguson"
               ),"michael mineo","sandra bland","sean bell","tamir rice"
            )
         ),
         (SENT,
            (OR,"bystander","bystanders","innocent bystander","innocent bystanders"
            ),

(OR,"death","deaths","die","died","kill","killed","killing","kills","shot","shooting",
"shootings"
            ),
            (OR,"by a cop","by cops","by police","undercover New York
City","undercover officer","undercover officers","undercover police","undercover"
            )
         ),
         (SENT,"Brussels","orgy","police"
         )
      ),
      (NOT,
         (OR,"correctional facility","Rikers"
         )
      )
   )
)
```

A.2 Attacks on Police

```
(OR,
   (OR,"assassination of police","assassinations of police","assata shakur","assaults
on police","assaulting a police","assaulting police","attack on police","attack on a
police","attacks on police","attacks on a police","battery on a police
officer","battery of a police officer","battle with the police","clash between police
and","confrontation with the police","cop shot","cop-shot","fallen police
officer","fallen police officers","fallen officer","fallen officers","ismaaiyl
brinsley","ismaaiyl a. brinsley","ismaaiyl abdullah brinsley","joanne
chesimard","joanne d. chesimard","john patrick bedell","killing of police","killings
of police","murder of a police","police fatalities","near-fatal attack on a
police","near-fatal attacks on police","officer slain","officers slain","killed a
police","killed police","killing a police","killing police","killing of new york
police officers","killing of police","murdered a police","murdered police","murdering
a police","shot a police","shot police","shot a state trooper","shot an officer","shot
the deputy","slain officer","slain officers","stab a police","stabs a police","stabbed
a police","stabbing a police","war on cops","war on police"
   ),
   (SENT,
      (OR,"cop","cops","deputy","deputies",
         (AND,
            (OR,"officer","officers"
            ),
            (NOT,"corrections"
            )
         ),"police","policemen","policeman","policewoman","policewomen","state
trooper","state troopers","sheriff","sheriffs",
         (OR,"l.a.p.d.","lapd"
         ),
         (OR,"n.y.p.d.","nypd"
         )
      ),
      (OR,"been shot","died in the line of duty","killed in the line of
duty","returned fire","were attacked","were critically injured","were injured","were
shot"
      )
   ),
   (AND,"police",
      (OR,"assaults the lieutenant","assaults the lieutenant","assaulted the
lieutenants","assaulted the lieutenant","assault officer","assaulted
officer","assaulting officer","assault officers","assaulted officers","assaulting
officers","attack in dallas against police officers","attacked officer","attack
officer","attacking officer","attacked officers","attack officers","attacking
officers","killed officers","killing of officers","kills officers","kill
officers","officer deaths","officers are killed","officers are being killed","officer
was bleeding","officers were fatally shot","officers are killed","officers were
killed","officers were shot","stabbed officer","stabbing officer","stabbing
```

```
officers","stab officer","trooper's killer","trooper&#8217;s killer","trooper
executed","trooper was killed","troopers were killed"
      )
   ),
   (ORDDIST_3,
      (NOTIN,

(OR,"assaulted","assaulting","attacked","attacking","killed","killing","stabbed","stab
bing","shot","shooting"
         ),
         (OR,"assaulted by","assaulting by","attacked by","attacking by","killed
by","killing by","stabbed by","stabbing by","shot by","shooting by"
         )
      ),
      (OR,"officer","officers","police","policeman","policewoman","state trooper"
      )
   ),
   (MIN_2,
      (OR,"ismaaiyl brinsley","ismaaiyl a. brinsley","ismaaiyl abdullah brinsley"
      ),"rafael ramos",
      (OR,"wen jian liu","wenjian liu"
      )
   ))
```

A.3 Philanthropy

Contains examples of escaped characters "$_L" for dollar sign.

```
(OR,
   (OR,
      _headline:"charity",
      _headline:"charities",
      _headline:"charitable",
      _headline:"donates $_L",
      _headline:"philanthropy",
      _headline:"philanthropies",
      _headline:"philanthropic",
      _nyt
      _kicker:"neediest cases","neediest cases"
   ),
   (OR,
      _meta\id="title":"donates $_L",
      _meta\id="description":"donates $_L"
   ),
   (SENT,
      _meta\id="description":"donates",
      _meta\id="description":"million"
```

```
   ),
   (AND,
      (OR,"challenge
grant","donate","donated","donates","donating","gift","gifts","pledge","foundation","f
und-raiser","fund-raiser","fundraisers","fund-raisers","fundraising","fund-raising","g
oodwill","ice bucket challenge","ice-bucket challenge","neediest cases","raise
money","raising money","raised money","raises money","telethon","telethons"
      ),
      (OR,"als","a.l.s.","amyotrophic lateral
sclerosis","charity","charities","endowment","endowments","Lou Gehrig's disease","Lou
Gehrig&#8217;s
disease","nonprofit","nonprofits","non-profit","non-profits","scholarship","scholarshi
ps"
      ),
      (NOT,
         (OR,"political party","political parties"
         )
      )
   ),
   (MINOC_2,"anonymous donor","anonymous donors","anonymous donation","anonymous
donations","philanthropist","philanthropists","philanthropy","charity","charities","ch
aritable","ford foundation","goodwill","grant-making","making grants","pledge
money","pledge money","rockefeller foundation","bill and melinda gates
foundation","salvation army","reynolds foundation","501c(3)","social
entrepreneur","social entrepreneurs","unreasonable institute",
      (SENT,
         (OR,"$_L"
         ),
         (OR,"challenge
grant","donation","donations","donated","donating","donates","donate","pledge"
         )
      )
   ),
   (MINOC_2,"crowdfunding","crowd
funding","crowd-funding","crowdtilt","gambitious","kickstarter","zokos"
   )
)
```

A.4 Shooting (Sport)

Contains examples of referencing other rules (NOT _tmac Murders or Basketball…)

```
(OR,"competitive shooter","competitive shooters","competitive shooting",
   (AND,
      (AND,
         (OR,"National Shooting Sports Foundation","shooting"
         ),
         (OR,"olympics","olympic","sport","sports"
         )
```

```
      ),
      (NOT,
         (OR,
            _tmac:"@Basketball",
            _tmac:"@Basketball (College)","death","deaths","Elle
Macpherson","football","goal","goals","goalie","golf","golfer","golfers","golfing","hu
nting","hunter","hunters",
            _tmac:"@Murders, Attempted Murders and Homicides","n.c.a.a.","soccer"
         )
      )
   ),
   (AND,
      (MINOC_3,"shooting","shooter","shooters"
      ),
      (MIN_2,"clay pigeons","controlled pull-away method","dover furnace","proper
stance","quartering targets","shooting grounds","sporting clays"
      )
   )
)
```

A.5 Astronomy

Relatively simple rule; contains examples of capitalization, wildcards, stemming (both POS-based and general), referencing other rules, referencing lists

```
(AND,

(MIN_2,"astronomy","astrophysic*","Osservatorio_C","celestial","lunar","universe","astronomical"
,"solar system@N","telescope@N","radiotelescope@","sky
imager","Telescopio_C","NASA_C","Infrared","Wavelength","universe@N","galaxy@N",
    (NOTIN,"observatory@N","Naval Observatory_C"
    )
  ),
  (MINOC_2,"Osservatorio_C",
    (NOTIN,
      (OR,"astronomy","astrophysic*","Observatory_C"
      ),
      (OR,"[Degrees]","Naval Observatory_C"
      )
    ),
    (SENT,
      (OR,"celestial","lunar","universe","astronomical","solar system@N"
      ),
      (OR,"methodical
observation@N","physics","chemistry","evolution","meterology","motion@N",(NOTIN,"science@
N","[Degrees]")
```

```
        ),
        (OR,"material object@N","phenomena"
        )
      ),
      (SENT,
        (OR,"universe@N","galaxy@N"
        ),
        (OR,"form@","formation@N","develop@","development@N"
        )
      ),
      (SENT,
        (OR,"observatory@N_C","telescope@N","radiotelescope@","sky imager","Telescopio_C"
        ),

(OR,"NASA_C","Infrared","Solar","Optical","radio","Mount_C","Mt._C","Mauna_C","Peak_C","Mir
ror","Wavelength","large","largest","UKIRT_C","WYIN","Galileo","Naval_C","MMT_C","meter","K
eck"
        )
      )
    ),
    (NOT,
      (OR,"toys","film festival","Galaxy Tab_C",
        _Title:"Dean's list",
        _Title:"Deans list",
        _Title:"Honor roll",
        _HeadLine:"Dean's list",
        _HeadLine:"Deans list",
        _HeadLine:"Honor roll",
        _tmac:"@Top/Disabled categories/Universal NOT",
        _tmac:"@Top/Disabled categories/Business content",
        _tmac:"@Top/Arts and entertainment/Entertainment/Movies",
        _tmac:"@Top/Environment and nature/Environment/Energy and the
environment/Alternative and sustainable energy/Solar power"
      )
    )
)
```

A.6 Celebrity

*More complex rule; contains examples of field targeting, capitalization, wildcards, stemming (both POS-based and general), positional markers, refere*CELEB_C*ncing other rules, referencing lists*

```
(AND,
  (OR,
    _Fixture:"^People$",
    _Fixture:"Celebrity Birthdays",
    _SlugLine:"Celebrity Birthdays",
    _Title:"Celebrity Birthdays",
    _SlugLine:"celeb-birthdays",
    _Title:"celeb-birthdays",
    _Title:"",
    _SlugLine:"CELEB_C",
    _Title:"^People-_C",
    _Title:"^US-People-_C",
    _Title:"^US--People-_C",
    _Title:"^PEOPLE-_C",
    _Title:"^US-PEOPLE-_C",
    _Title:"^US--PEOPLE-_C",
    _Title:"-People-_L",
    _Title:"-PEOPLE-_L",
    _FilingSubSubject:"Celebrities",
    _FilingSubSubject:"Celeb",
    (AND,
      (OR,
        _Title:"A&E_C",
        _Title:"A&amp;E_C"
      ),
      (OR,
        _Title:"CEL_C"
      )
    ),
    (AND,
      (OR,
        (NOTIN,
          (OR,
            _Title:"^People_C",
            _Title:"^PEOPLE_C"
          ),
          (OR,"People's","People Mag*"
          )
        )
      ),
      (OR,"[PROFESSIONAL_ATHLETE]","[OLYMPIC_ATHLETE]","[FASHION_DESIGNER]",
        (DIST_8,
          (NOTIN,"[All athletes]",
```

```
            (OR,"[PROFESSIONAL_ATHLETE]","[OLYMPIC_ATHLETE]"
            )
          ),
          (OR,"[Sports names]","[Sports roles]"
          )
        ),
        (NOTIN,
          (OR,"[ENTERTAINMENT_FIGURE]","[Celebrity]"
          ),
          (OR,"[Ambiguous entertainment figures]","[Celebrity metaphor]","like a star","like
stars","star for the day","like a celebrity","like celebrities","star witness@N","star-witness*"
          )
        )
      )
    ),
    (MIN_2,
      (NOTIN,
        (OR,"celebrity@N","celeb","celebs"
        ),
        (OR,"[Celebrity metaphor]","like a celebrity","like celebrities",
          _KeywordLine:"celeb*"
        )
      ),
      (OR,"red carpet","red-carpet"
      ),
      (OR,
        _Title:"[ENTERTAINMENT_FIGURE]",
        _HeadLine:"[ENTERTAINMENT_FIGURE]",
        _SlugLine:"[ENTERTAINMENT_FIGURE]"
      ),
      (SENT,
        (DIST_12,
          (OR,"[Narcotics]","[Pharmaceuticals]","[Medical Conditions]","[Religions]","red
carpet","red-carpet","neighborhood","neighbor@N","guest judge","famous",
            (NOTIN,
              (OR,"fame","gossip@","[Celebrity news]"
              ),
              (OR,"gossip girl","The Fame_C","jane's addiction","janes addiction"
              )
            ),"of the year","nominated","presidential aspirations"
          ),
          (OR,
            (NOTIN,"star@N",
```

```
                (OR,"all star@N","all-star*","like a star","like stars","star for the day","star
witness@N","star-witness*","[Celebrity metaphor]"
                )

),"superstar@N","starlet@N","socialite@N","celebutant*","supermodel@N","[FASHION_DESIG
NER]","[PROFESSIONAL_ATHLETE]","[OLYMPIC_ATHLETE]",
            (DIST_8,
              (NOTIN,"[All athletes]",
                (OR,"[PROFESSIONAL_ATHLETE]","[OLYMPIC_ATHLETE]"
                )
              ),
              (OR,"[Sports names]","[Sports roles]"
              )
            ),
            (NOTIN,"[ENTERTAINMENT_FIGURE]",
              (OR,"[Ambiguous entertainment figures]","[Celebrity roles]","[Celebrity metaphor]",
                (ORDDIST_4,
                  (OR,"like"
                  ),
                  (OR,"[ENTERTAINMENT_FIGURE]"
                  )
                )
              )
            )
          )
        )
      ),
      (MINOC_3,
        (NOTIN,
          (OR,"celebrity@N","celeb","celebs"
          ),
          (OR,
            _KeywordLine:"celeb*"
          )
        ),
        (SENT,
          (OR,"[Narcotics]","[Pharmaceuticals]","[Medical Conditions]","[Religions]","red
carpet","red-carpet","neighborhood","neighbor@N","guest judge","famous",
            (NOTIN,
              (OR,"[Criminal trials]","fame","gossip@","[Celebrity news]"
              ),
              (OR,"time trial@N","gossip girl","The Fame_C","jane's addiction","janes addiction"
```

```
            )
          )
        ),
        (OR,
          (NOTIN,"star@N",
            (OR,"all star@N","all-star*","like a star","like stars","star for the day","star
witness@N","star-witness*","[Celebrity metaphor]"
            )

),"socialite@N","superstar@N","starlet@N","celebutant*","supermodel@N","[FASHION_DESIG
NER]",
          (NOTINSENT,
            (NOTIN,"[ENTERTAINMENT_FIGURE]",
              (OR,"[Ambiguous entertainment figures]","[Celebrity roles]","[Celebrity metaphor]"
              )
            ),
            (OR,"plays a","played","playing","film about","movie about","role@N","starring"
            )
          )
        )
      ),
      (SENT,
        (OR,"[Narcotics]","[Pharmaceuticals]","[Religions]","red
carpet","red-carpet","neighborhood","neighbor@N","guest judge","famous",
          (NOTIN,
            (OR,"[Criminal trials]","fame","gossip@","[Celebrity news]"
            ),
            (OR,"time trial@N","gossip girl","The Fame_C","jane's addiction","janes addiction"
            )
          )
        ),
        (OR,"[PROFESSIONAL_ATHLETE]","[OLYMPIC_ATHLETE]",
          (DIST_8,
            (NOTIN,"[All athletes]",
              (OR,"[PROFESSIONAL_ATHLETE]","[OLYMPIC_ATHLETE]"
              )
            ),
            (OR,"[Sports names]","[Sports roles]"
            )
          )
        )
      )
    ),
```

```
    (AND,
      (OR,
        _tmac:"@Top/Arts and entertainment/Entertainment/Movies/Movie premieres",
        _tmac:"@Top/Arts and entertainment/Entertainment/Award shows",
        _tmac:"@Top/Events/Entertainment events/Academy of Country Music Awards",
        _tmac:"@Top/Events/Entertainment events/Emmy Awards",
        _tmac:"@Top/Events/Entertainment events/Golden Globe Awards",
        _tmac:"@Top/Events/Entertainment events/Grammy Awards",
        _tmac:"@Top/Events/Entertainment events/People's Choice Awards",
        _tmac:"@Top/Events/Entertainment events/Screen Actors Guild Awards",
        _tmac:"@Top/Events/Entertainment events/Tony Awards",
        _tmac:"@Top/Events/Entertainment events/Academy Awards"
      ),
      (SENT,
        (OR,"[Celebrity events]"
        ),
        (OR,"supermodel@N","[FASHION_DESIGNER]","[PROFESSIONAL_ATHLETE]",
          (NOTIN,"[All athletes]","[PROFESSIONAL_ATHLETE]"
          ),
          (NOTIN,
            (OR,"[Celebrity]","[ENTERTAINMENT_FIGURE]"
            ),
            (OR,"[Ambiguous entertainment figures]","star
witness@N","star-witness*","[Celebrity metaphor]"
            )
          )
        )
      )
    ),
    (AND,
      (OR,
        _MediaType:"Photo",
        (AND,
          _MediaType:"Video","video contains ONLY natural sound"
        )
      ),
      (OR,
        _Title:"[ENTERTAINMENT_FIGURE]",
        _HeadLine:"[ENTERTAINMENT_FIGURE]",
        _SlugLine:"[ENTERTAINMENT_FIGURE]",
        _NameLine:"[ENTERTAINMENT_FIGURE]",
        _OverLine:"[ENTERTAINMENT_FIGURE]",
        (ORDDIST_6,"Pictured:_L","[ENTERTAINMENT_FIGURE]"
```

```
        ),
        (SENT,
          (OR,"[CELEBRITY_CHEF]"
          ),

(OR,"[Celebrity]","Chef@N","Cook@","Cookbook@N","Cuisin*","Culinar*","Restaurant@N","Res
taurateur@N","[Celebrity events]",
            (ORDDIST_3,"[CELEBRITY_CHEF]","pose@N"
            )
          )
        ),
        (SENT,
          (OR,"[COMEDIAN]"
          ),

(OR,"[Celebrity]","comedian@N","Film@","Hollywood","Movie@N","Theater","Theatre","publicist
@N","paparazz*","[Celebrity events]",
            (ORDDIST_3,"[COMEDIAN]","pose@N"
            )
          )
        ),
        (SENT,
          (OR,"[DIRECTOR]"
          ),

(OR,"[Celebrity]","Director@N","directing","directed","Movie@N","Film@","Hollywood","Studio","
Producer@","box-office","box office","DVD_C","[Celebrity events]",
            (ORDDIST_3,"[DIRECTOR]","pose@N"
            )
          )
        ),
        (SENT,
          (OR,"[MOVIE_ACTOR]"
          ),

(OR,"[Celebrity]","Actor","Actress","Film@","Hollywood","Movie@N","Theater","Theatre","Tony
Award*_C","Oscar_C","Oscars_C","Academy Award*_C","publicist@N","paparazz*","[Celebrity
events]",
            (ORDDIST_3,"[MOVIE_ACTOR]","pose@N"
            )
          )
        ),
        (SENT,
```

```
            (OR,"[MISC_ENTERTAINER]","[TV_PERSONALITY]"
            ),
            (OR,"[Celebrity]","entertainer","TV
personality","Actor","Actress","Film@","Hollywood","Movie@N","Theater","Theatre","publicist@N
","paparazz*","[Celebrity events]",
              (ORDDIST_3,
                (OR,"[MISC_ENTERTAINER]","[TV_PERSONALITY]"
                ),"pose@N"
              )
            )
          ),
          (SENT,
            (OR,"[MUSICIAN]"
            ),

(OR,"[Celebrity]","Billboard_C","Composer","Concert","Music","Musical*","Musician","Opera","ro
ck star","Singer","Singer-songwriter","Songwriter","soundtrack","Stardom","Superstar","recording
artist@N","paparazz*","[Music genres]","[Celebrity events]",
              (ORDDIST_3,"[MUSICIAN]","pose@N"
              )
            )
          ),
          (SENT,
            (DIST_10,
              (OR,"[MODEL]","[FASHION_DESIGNER]"
              ),
              (OR,"[Celebrity]","Model","supermodel","Modeling","Fashion","designer","[Celebrity
events]",
                (ORDDIST_3,
                  (OR,"[MODEL]","[FASHION_DESIGNER]"
                  ),"pose@N"
                )
              )
            )
          )
        )
      )
    ),
    (NOT,
      (OR,"Celebrity Cruise_C","Celebrity Eclipse_C","Celebrity ships","Weddings by Martha
Stewart_C","Music City Star_C",
        (AND,
          (OR,
```

```
          _Source:"Star",
          _Provider:"Star"
        ),
        (NOT,"[ENTERTAINMENT_FIGURE]"
        )
      ),
      _Title:"Travel-Trip",
      (AND,
        _Title:"FEA_C",
        _Title:"Travel_C"
      ),
      (MINOC_2,"box office@N","boxoffice*",
        (DIST_10,
          (OR,"tickets"
          ),
          (OR,"on sale","discount*"
          )
        ),
        (SENT,
          (OR,"admission"
          ),
          (OR,"discount*","children"
          )
        ),"ticket sales","reserved seating","ticket information"
      ),
      (ORDDIST_3,"Rated_C",
        (OR,"R_C","PG_C","PG-13_C","G_C","NC-17_C"
        )
      ),
      _tmac:"@Top/Disabled categories/Universal NOT",
      _tmac:"@Top/Disabled categories/Business content",
      _tmac:"@Top/Disabled categories/Reviews/Dance review",
      _tmac:"@Top/Disabled categories/Reviews/Movie review",
      _tmac:"@Top/Disabled categories/Reviews/Music review",
      _tmac:"@Top/Disabled categories/Reviews/Theater review",
      _tmac:"@Top/Disabled categories/Reviews/TV review",
      _tmac:"@Top/Sports/Athlete health",
      _tmac:"@Top/Sports/Athlete health/Athlete injuries"
    )
  )
)
```

A.7 Hurricanes

More complex rule; contains examples of field targeting, capitalization, wildcards, stemming (both POS-based and general), positional markers, referencing other rules, referencing lists

```
(AND,
  (OR,
    (NOTINDIST_6,
      (NOTIN,"hurricane@N_C",
        (OR,"Carolina Hurricanes_C","Miami Hurricanes_C","Hurricane Cent*_C","hurricane
shelter@N","hurricane-strength","hurricane strength","hurricane-proof*","hurricane
proof*","hurricane-protection","hurricane protection","hurricane hazard response","Hurricane
Insurance","Hurricane Point_C","Hurricane Street","Hurricane Ave*_C","Hurricane
Road","Hurricane Drive","Hurricane Ridge_C"
        )
      ),
      (OR,"just shy of","years ago","not an actual","potential"
      )
    ),"Superstorm Sandy",
    (ORDDIST_1,"Hurricane@N_C",

(OR,"[AtlanticCycloneNames]","[CentNorthPacificCycloneNames]","[EastNorthPacificCycloneNa
mes]"
      )
    )
  ),
  (OR,
    (AND,
      (OR,
        (MINOC_5,
          (NOTINDIST_6,
            (NOTIN,"hurricane@N",
              (OR,"Carolina Hurricanes_C","Miami Hurricanes_C","Hurricane
Cent*_C","hurricane shelter@N","hurricane-strength","hurricane
strength","hurricane-proof*","hurricane proof*","hurricane-protection","hurricane
protection","hurricane hazard response","Hurricane Insurance","Hurricane Point_C","Hurricane
Street","Hurricane Ave*_C","Hurricane Road","Hurricane Drive","Hurricane Ridge_C"
              )
            ),
            (OR,"just shy of","years ago","not an actual","potential"
            )
          ),"Superstorm Sandy"
```

```
        ),
        (NOTIN,
          (OR,
            _HeadLine:"superstorm",
            _HeadLine:"hurricane",
            _Title:"hurricane",
            _Title:"superstorm",
            _ExtendedHeadLine:"hurricane",
            _ExtendedHeadLine:"superstorm",
            _ExtendedHeadLine:"frankenstorm",
            _HeadLine:"frankenstorm",
            _Title:"frankenstorm"
          ),
          (OR,
            _ExtendedHeadLine:"Hurricane Cent*_C",
            _HeadLine:"Hurricane Cent*_C",
            _Title:"Hurricane Cent*_C",
            _ExtendedHeadLine:"Hurricane insurance",
            _HeadLine:"Hurricane insurance",
            _Title:"Hurricane insurance"
          )
        )
      ),
      (MINOC_3,
        (NOTINDIST_4,
          (OR,"killing","killed","dead"
          ),
          (OR,"vegetation","grass@N","plants","brush","vehicle@N","wreck@N","crash"
          )

),"missing","injury@","fatality@","injure@","death@","damage","storm@N","victims","survivors","
devastation","devastate@","trajectory","levee@",
        (ORDDIST_2,"category",
          (OR,"[0-10 Spelled]","[Digits]"
          )
        ),"Corps of Engineers","flooding","wind@N","deathtoll",
        (ORDDIST_3,
          (OR,"issue@","hurricane"
          ),"warning@N"
        ),
        (ORDDIST_2,
          (OR,"coastal","low-lying","low lying"
          ),
```

```
        (OR,"town@","village@","area@N","community@N","city@N"
        )
      ),"struck","strikes",
      (ORDDIST_2,
        (OR,"come@","crash@","slam@"
        ),"ashore"
      ),"lashing","surging","surge","landfall","lumber@","land-fall",
      (ORDDIST_3,"reach@",
        (OR,"coast","shore"
        )

),"evacuate@","evacuation@N","evacuee*","intensity@","displaced","disaster@N","superstorm",
"frankenstorm","without power","power outage@N","gas shortage@"
      )
    ),
    (AND,
      (OR,
        (AND,
          (ORDDIST_1,"Hurricane_C",

(OR,"[AtlanticCycloneNames]","[CentNorthPacificCycloneNames]","[EastNorthPacificCycloneNa
mes]"
            )
          ),
          (MINOC_2,
            (NOTINDIST_6,
              (NOTIN,"hurricane@N",
                (OR,"Carolina Hurricanes_C","Miami Hurricanes_C","Hurricane
Cent*_C","hurricane shelter@N","hurricane-strength","hurricane
strength","hurricane-proof*","hurricane proof*","hurricane-protection","hurricane
protection","hurricane hazard response","Hurricane Insurance","Hurricane Point_C","Hurricane
Street","Hurricane Ave*_C","Hurricane Road","Hurricane Drive","Hurricane Ridge_C"
                )
              ),
              (OR,"just shy of","years ago","not an actual","potential"
              )
            ),"Superstorm Sandy","frankenstorm"
          )
        ),
        (MINOC_3,
          (NOTINDIST_6,
            (NOTIN,"hurricane@N",
```

```
                (OR,"Carolina Hurricanes_C","Miami Hurricanes_C","Hurricane
Cent*_C","hurricane shelter@N","hurricane-strength","hurricane
strength","hurricane-proof*","hurricane proof*","hurricane-protection","hurricane
protection","hurricane hazard response","Hurricane Insurance","Hurricane Point_C","Hurricane
Street","Hurricane Ave*_C","Hurricane Road","Hurricane Drive","Hurricane Ridge_C"
                )
              ),
              (OR,"just shy of","years ago","not an actual","potential"
              )
            ),"Superstorm Sandy","frankenstorm"
          )
        ),
        (MIN_4,
          (NOTINDIST_4,
            (OR,"killing","killed","dead"
            ),
            (OR,"vegetation","grass@N","plants","brush","vehicle@N","wreck@N","crash"
            )

),"missing","injury@","fatality@","injure@","death@","damage","storm@N","victims","survivors","
devastation","devastate@","trajectory","levee@",
          (ORDDIST_2,"category",
            (OR,"[0-10 Spelled]","[Digits]"
            )
          ),"Corps of Engineers","flood*","wind@N","deathtoll",
          (ORDDIST_3,
            (OR,"issue@","hurricane"
            ),"warning@N"
          ),
          (ORDDIST_4,
            (OR,"coastal","low-lying","low lying"
            ),
            (OR,"town@","village@","area@N","community@N","city@N"
            )
          ),"struck","strikes",
          (ORDDIST_2,
            (OR,"come@","crash@","slam@"
            ),"ashore"
          ),"lashing","surging","surge","landfall","lumber@","land-fall",
          (ORDDIST_3,"reach@",
            (OR,"coast","shore"
            )
```

```
),"evacuate@","evacuation@N","evacuee*","intensity@","displaced","disaster@N","superstorm",
"frankenstorm","without power","power outage@","Gas shortage@","rescue@"
        )
      ),
      (AND,
        _MediaType:"photo",
        (OR,
          (ORDDIST_1,"Hurricane@N_C",

(OR,"[AtlanticCycloneNames]","[CentNorthPacificCycloneNames]","[EastNorthPacificCycloneNa
mes]"
            )
          ),
          (NOTIN,
            (OR,
              _ExtendedHeadLine:"hurricane",
              _HeadLine:"hurricane",
              _Title:"hurricane"
            ),
            (OR,
              _ExtendedHeadLine:"Hurricane Cent*_C",
              _HeadLine:"Hurricane Cent*_C",
              _Title:"Hurricane Cent*_C",
              _ExtendedHeadLine:"Hurricane insurance",
              _HeadLine:"Hurricane insurance",
              _Title:"Hurricane insurance"
            )
          ),
          (MINOC_2,
            (NOTINDIST_6,
              (NOTIN,"hurricane@N",
                (OR,"Carolina Hurricanes_C","Miami Hurricanes_C","Hurricane
Cent*_C","hurricane shelter@N","hurricane-strength","hurricane
strength","hurricane-proof*","hurricane proof*","hurricane-protection","hurricane
protection","hurricane hazard response","Hurricane Insurance","Hurricane Point_C","Hurricane
Street","Hurricane Ave*_C","Hurricane Road","Hurricane Drive","Hurricane Ridge_C"
                )
              ),
              (OR,"just shy of","years ago","not an actual","potential"
              )
            )
          ),"Superstorm Sandy"
```

```
        )
      )
    ),
    (NOT,
      (OR,
        _SlugLine:"Crop report",
        _SlugLine:"Forecast",
        _SlugLine:"*-forecast",
        _SlugLine:"*-HKN-*",
        _SlugLine:"*weather outlook*",
        _SlugLine:"warning",
        _tmac:"@Top/Disabled categories/Universal NOT",
        _tmac:"@Top/General news/Weather/Weather forecasts",
        _Title:"Press Release",
        _Title:"FD-(FAIR-DISCLOSURE)_L",
        _HeadLine:"Press Release",
        _HeadLine:"FD-(FAIR-DISCLOSURE)_L",
        (MINOC_3,"[Sports Terms]","[Sports roles]","[Sports positions]","CQ
Transcriptions_C","Lady Dolphins_C"
        ),
        _DateLine:"Hurricane",
        (AND,
          _Source:"The Spectrum_C",
          (OR,"HURRICANE_C","St. George Spectrum"
          )
        ),
        (MINOC_4,
          (ORDDIST_1,"climate",

(OR,"change","model@N","variation@N","indicator@N","system@N","science","scientist@N"
              )
          ),"global warming"
        ),
        _tmac:"@Top/Arts and entertainment/Entertainment/Books and literature",
        _tmac:"@Top/Arts and entertainment/Entertainment/Books and literature/Fiction"
      )
    ))
```

A.8 Sample Grammar Rule

```
#ROOT=*ABOUT
```

*ORGTYPE = Foundation
*ORGTYPE = FOUNDATION
*ORGTYPE = Association
*ORGTYPE = ASSOCIATION
*ORGTYPE = Associated
*ORGTYPE = ASSOCIATED
*ORGTYPE = University
*ORGTYPE = UNIVERSITY
*ORGTYPE = College
*ORGTYPE = COLLEGE
*ORGTYPE = Endowment
*ORGTYPE = ENDOWMENT
*ORGTYPE = Trust
*ORGTYPE = TRUST

*ORG = !Company
*ORG = *CAP *ORGTYPE
*ORG = *CAP , *ORGTYPE
*ORG = #cap *ORGTYPE
*ORG = #cap , *ORGTYPE
*ORG = and *ORGTYPE
*ORG = And *ORGTYPE
*ORG = AND *ORGTYPE
*ORG = & *ORGTYPE

*ABOUT = About *ORG
*ABOUT = ABOUT *ORG

*CAP = !InitialCap
*CAP = !{ALL CAPS}

*W = !InitialCap
*W = !{ALL CAPS}
*W = !Lowercase

*ABOUT = About *CAP
*ABOUT = ABOUT *CAP
*ABOUT = About *W *ORG

```
*ABOUT = ABOUT *W *ORG
*ABOUT = About *W *W *ORG
*ABOUT = ABOUT *W *W *ORG
*ABOUT = About *W *W *W *ORG
*ABOUT = ABOUT *W *W *W *ORG
*ABOUT = About *W *W *W *W *ORG
*ABOUT = ABOUT *W *W *W *W *ORG


# *ABOUT = About #cap
# *ABOUT = ABOUT #cap
# *ABOUT = About #w *ORG
# *ABOUT = ABOUT #w *ORG
# *ABOUT = About #w #w *ORG
# *ABOUT = ABOUT #w #w *ORG
# *ABOUT = About #w #w #w *ORG
# *ABOUT = ABOUT #w #w #w *ORG
# *ABOUT = About #w #w #w #w *ORG
# *ABOUT = ABOUT #w #w #w #w *ORG
```

A.9 ‘Bylines’ Grammar Rule

```
#ROOT=*BY

*REP = reporter
*REP = Reporter
*REP = journalist
*REP = Journalist
*REP = editor
*REP = Editor
*REP = editor-in-chief
*REP = Editor-in-Chief
*REP = publisher
*REP = Publisher

*MID = is a
*MID = is an
*MID = was a
*MID = was an

*NAME = !Person
```

*NAME = !{Fullname female}
*NAME = !{Fullname male}
*NAME = !{Ambiguous names}

*BY = *NAME *MID !Newspapers *REP
*BY = *NAME *MID #w !Newspapers *REP
*BY = *NAME *MID *REP for !Newspapers
*BY = *NAME *MID #w *REP for !Newspapers
*BY = *NAME *MID *REP with !Newspapers
*BY = *NAME *MID #w *REP with !Newspapers
*BY = *NAME *MID *REP for the !Newspapers
*BY = *NAME *MID #w *REP for the !Newspapers
*BY = *NAME is *REP of !Newspapers
*BY = *NAME is *REP for !Newspapers
*BY = *NAME is *REP with !Newspapers
*BY = *NAME is *REP of the !Newspapers
*BY = *NAME is *REP for the !Newspapers
*BY = *NAME is *REP with the !Newspapers

*BY = *NAME *MID !{News agencies} *REP
*BY = *NAME *MID #w !{News agencies} *REP
*BY = *NAME *MID *REP for !{News agencies}
*BY = *NAME *MID #w *REP for !{News agencies}
*BY = *NAME *MID *REP with !{News agencies}
*BY = *NAME *MID #w *REP with !{News agencies}
*BY = *NAME *MID *REP for the !{News agencies}
*BY = *NAME *MID #w *REP for the !{News agencies}
*BY = *NAME is *REP of !{News agencies}
*BY = *NAME is *REP for !{News agencies}
*BY = *NAME is *REP with !{News agencies}
*BY = *NAME is *REP of the !{News agencies}
*BY = *NAME is *REP for the !{News agencies}
*BY = *NAME is *REP with the !{News agencies}

Appendix B: Hit Highlighting Examples

B.1 An Example Rule

To help illustrate how hit highlighting is meant to work, we will use this example rule (in XQuery FT syntax).

```
(
      (
       {("anemia.*", "anaemia.*")} any word occurs at least 4 times using
wildcards
      )
      ftor
      (
       (
         {("anemia.*", "anaemia.*")} any word occurs at least 3 times using
wildcards
       )
       ftand
       (
          {("patient", "anemic",
"treatment","AOP","haemoglobin","hemoglobin","red blood cell","red
blood-cell","thalassaemias","sickle-cell","sickle
cell","RBC","hypoxia","symptom","hematocrit","Microcytic","Normocytic","Macrocy
tic","erythropoietin","anaemic","aplastic","hemolytic","disorder","disease","my
elodysplastic","illness")} any word occurs at least 4 times using stemming
       )

     )
   )
```

B.2 An Example Document

Here is a document (in XML) which does match the “anemia” rule:

```
<Publication Version="5.3.1" xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns="http://ap.org/schemas/03/2005/appl"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <HeadLine>Calling attention to dangers of sickle cell anemia</HeadLine>
    <body.content>
        <block>
            <p>September is National Sickle Cell Awareness Month.  First
officially recognized by the federal government in 1983, observance calls
attention to the genetic disease that affects African Americans the most, with
about 1 in 365 African-American children born with the ailment.</p>
```

Sickle cell anemia is a blood disorder that’s inherited — meaning it’s passed down from parents to their children. Babies are born with the disease when they inherit two abnormal genes (one from each parent). These genes cause the body’s red blood cells to change shape.

It’s a chronic disorder causing pain in the extremities and back, infections, organ failure and other tissue damage, skin infections, iron deficiency, loss of eyesight, severe blood clots and strokes.

Normal red blood cells are shaped like discs or donuts with the centers partly scooped out. They are soft and flexible so they can easily move through very small blood vessels and deliver oxygen throughout a person’s body.

Sickle cells are stiffer. Instead of being disc-shaped, they’re curved like crescent moons, or an old farm tool known as a sickle. That’s where the disease gets its name.

While sickle cell disease primarily affects persons of African descent, it is also present in Portuguese, Spanish, French Corsicans, Sardinians, Sicilians, mainland Italians, Greeks, Turks and Cypriots. Sickle cell disease also appears in Middle Eastern countries and Asia.

The year 1910 is regarded as the date of the discovery of the disease.

The disorder, often abbreviated as SCD, had been present in Africa for at least 5,000 years and has been known by many names in many tribal languages. What we call its discovery in 1910 occurred, not in Africa, but in the United States. A young man named Walter Clement Noel from the island of Grenada, a dental student studying in Chicago, went to Dr. James B. Herrick with complaints of pain episodes and symptoms of anemia. Herrick was a cardiologist and not too interested in Noel’s case so he assigned a resident, Dr. Ernest Irons, to the case. Irons examined Noel’s blood under the microscope and saw red blood cells he described as “having the shape of a sickle.” When Herrick saw this in the chart, he became interested because he saw that this might be a new, unknown disease. He subsequently published a paper in one of the medical journals in which he used the term “sickle shaped cells.”

Originally from Africa and brought to the Americas by the forced immigration of slaves, it is more frequent where the proportion of African descendants is greater. Carriers of the sickle cell trait have some resistance to the often-fatal malaria. This is why it is found more frequently in persons of Middle Eastern, Indian, Mediterranean and African heritage because those geographic regions are most prone to malaria.

However, in areas such as the U.S., where malaria is not a problem, the trait no longer provides a survival advantage. Instead, it poses a threat in children of carriers who inherit the sickle cell gene from both parents.

There are several types of sickle cell disease, and the most common one is sickle cell anemia. It makes you feel tired, because your blood cells don’t last as long as other people’s do.

Normally, red blood cells live about 120 days. New ones that replace them are made in the soft, spongy center of your bones called the marrow. If you have sickle cell anemia, your red blood cells start dying after

```
only 10 to 20 days. Your bone marrow can’t replace them fast enough, which
causes anemia. Red blood cells carry oxygen around your body, which gives you
energy. If you don’t have enough of them, you’ll get tired and you might also
feel short of breath.</p>
            <p>SCD patients learn to function in a constant state of pain and
when that pain becomes debilitating, they often end up in the emergency room.
What brings people to the emergency room is called a crisis, in which the blood
can’t deliver oxygen to the extremities. It is a sudden onset of severe,
excruciating pain, which people describe as feeling as if all their bones are
breaking. The pain is so bad and sudden that people require high doses of pain
killers.</p>
            <p>Contrary to the belief of many in the medical profession, there
is no data to support that sickle cell patients have any more likelihood of
being addicted than anybody else.</p>
            <p>Acute and unpredictable pain, severe enough to require large
doses of narcotics, is a well-documented feature of the disease. When many
patients arrive at emergency rooms, they often have great difficulty getting
the treatment they need.</p>
            <p>Racism and the disease stigma itself are two barriers that you
just can’t get away from. Clearly we can’t pretend that racism doesn’t play
some part in this. If this were a white disease, people still wouldn’t be dying
in their 40s. That’s the bottom line. Sickle cell was “discovered” 106 years
ago and there is only one drug, hydroxyurea, and blood transfusions to treat
it.</p>
            <p>There has been progress in sickle cell disease. People didn’t
used to live to be adults. Kids would die of stroke or of really bad infections
before they were 20, and in some countries they still do.</p>
        </block>
    </body.content>
</Publication>
```

B.3 Hit Highlighting Example: the Rule

To highlight how a document matches a rule, it is necessary to markup both the rule and the document with TEI “milestones”, which indicate which parts of the document match which parts of the rule. Note that a given part of a document could match more than one part of a rule.

The marked up rule:

```
(
     (<milestone xmlns="http://www.tei-c.org/ns/1.0" n="m1" xml:id="r1"
spanTo="#r2"/>(
       {"anemia.*", "anaemia.*")} any word occurs at least 4 times using
wildcards
     )<milestone xmlns="http://www.tei-c.org/ns/1.0" n="m2" xml:id="r2"/>
     ftor
     (
```

```
       <milestone xmlns="http://www.tei-c.org/ns/1.0" n="m3" xml:id="r3"
spanTo="#r4"/>(
         {("anemia.*", "anaemia.*")} any word occurs at least 3 times using
wildcards
       )<milestone xmlns="http://www.tei-c.org/ns/1.0" n="m4" xml:id="r4"/>
       ftand
       <milestone xmlns="http://www.tei-c.org/ns/1.0" n="m5" xml:id="r5"
spanTo="#r6"/>(
          {("patient", "anemic",
"treatment","AOP","haemoglobin","hemoglobin","red blood cell","red
blood-cell","thalassaemias","sickle-cell","sickle
cell","RBC","hypoxia","symptom","hematocrit","Microcytic","Normocytic","Macrocy
tic","erythropoietin","anaemic","aplastic","hemolytic","disorder","disease","my
elodysplastic","illness")} any word occurs at least 4 times using stemming
       )<milestone xmlns="http://www.tei-c.org/ns/1.0" n="m6" xml:id="r6"/>

     )
   )
```

The clauses of the rule have been identified by the `<milestone>` elements. Each clause is marked by a pair of start and end `<milestone>` elements. Each `<milestone>` element has a document-unique identifier (the `@xml:id` attribute). The first `<milestone>` element in each pair has a `@spantTo` attribute, which references the `@xml:id` of the second `<milestone>` element in the pair. (This is different than the way XML documents normally work: normally you would have `<milestone>` at the start of a clause and `</milestone>` to indicate the end. However, that only works when you have a strict, hierarchical tree. In our case, it is quite likely that we will wind up with overlapping phrases. Which is why we're using TEI's `<milestone>` technique instead).

As we shall see below, the `@n` attribute values of the milestone elements are referenced by `<milestone>`s in the hit highlight document.

B.4 Hit Highlighting Example: the Document

The document - which does match the above rule - is also marked up with `<milestone>` elements.

Again, as with the rule, the `<milestone>`s appear in pairs, with the first `<milestone>` referencing the second via the `@spanTo` attribute. The `@n` attribute in the `<milestone>` pairs is a reference to which part of the rule is matched, i.e. to the `@n` attribute of the corresponding rule `<milestone>`.

Note that some parts of the document may match more than one part of a rule. In this case, the document part will have a `<milestone>` pair for each matching part of the rule.

<?xml version="1.0" encoding="utf-8"?><Publication
xmlns="http://ap.org/schemas/03/2005/appl"
xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:xhtml="http://www.w3.org/1999/xhtml"
xmlns:date="http://exslt.org/dates-and-times"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" Version="5.3.1">
    <HeadLine>Calling attention to dangers of <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="m1" spanTo="#m2"/>sickle
cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="m2"/>
<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="m3"
spanTo="#m4"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="m5"
spanTo="#m6"/>anemia<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2"
xml:id="m6"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1"
xml:id="m4"/></HeadLine>
    <body.content>
        <block>
            September is National <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="m7" spanTo="#m8"/>Sickle
Cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="m8"/>
Awareness Month. First officially recognized by the federal government in
1983, observance calls attention to the genetic <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="m9"
spanTo="#m10"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="m10"/> that affects African Americans the most, with about 1 in 365
African-American children born with the ailment.
            <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="sd0e31" spanTo="#d0e31"/>Sickle cell<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e31"/> <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e34"
spanTo="#d0e34"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2"
xml:id="sd0e342" spanTo="#d0e342"/>anemia<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="sd0e342"/><milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e34"/><milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e34"/> is a blood
<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e37"
spanTo="#d0e37"/>disorder<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="d0e37"/> that’s inherited — meaning it’s passed down from parents to
their children. Babies are born with the <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e40"
spanTo="#d0e40"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="d0e40"/> when they inherit two abnormal genes (one from each parent).
These genes cause the body’s <milestone xmlns="http://www.tei-c.org/ns/1.0"
n="r5" xml:id="sd0e43" spanTo="#d0e43"/>red blood cells<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e43"/> to change
shape.
            It’s a chronic <milestone xmlns="http://www.tei-c.org/ns/1.0"
n="r5" xml:id="sd0e49" spanTo="#d0e49"/>disorder<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e49"/> causing pain in the

extremities and back, infections, organ failure and other tissue damage, skin infections, iron deficiency, loss of eyesight, severe blood clots and strokes.

Normal <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e55" spanTo="#d0e55"/>red blood cells<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e55"/> are shaped like discs or donuts with the centers partly scooped out. They are soft and flexible so they can easily move through very small blood vessels and deliver oxygen throughout a person's body.

<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e60" spanTo="#d0e60"/>Sickle cells<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e60"/> are stiffer. Instead of being disc-shaped, they're curved like crescent moons, or an old farm tool known as a sickle. That's where the <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e63" spanTo="#d0e63"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e63"/> gets its name.

While <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e69" spanTo="#d0e69"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e69"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e72" spanTo="#d0e72"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e72"/> primarily affects persons of African descent, it is also present in Portuguese, Spanish, French Corsicans, Sardinians, Sicilians, mainland Italians, Greeks, Turks and Cypriots. <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e75" spanTo="#d0e75"/>Sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e75"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e78" spanTo="#d0e78"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e78"/> also appears in Middle Eastern countries and Asia.

The year 1910 is regarded as the date of the discovery of the <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e84" spanTo="#d0e84"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e84"/>.

The <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e90" spanTo="#d0e90"/>disorder<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e90"/>, often abbreviated as SCD, had been present in Africa for at least 5,000 years and has been known by many names in many tribal languages. What we call its discovery in 1910 occurred, not in Africa, but in the United States. A young man named Walter Clement Noel from the island of Grenada, a dental student studying in Chicago, went to Dr. James B. Herrick with complaints of pain episodes and <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e93" spanTo="#d0e93"/>symptoms<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e93"/> of <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e96" spanTo="#d0e96"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e96" spanTo="#d0e96"/><milestone

xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="sd0e962" spanTo="#d0e962"/>anemia<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="sd0e962"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e96"/>. Herrick was a cardiologist and not too interested in Noel's case so he assigned a resident, Dr. Ernest Irons, to the case. Irons examined Noel's blood under the microscope and saw <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e99" spanTo="#d0e99"/>red blood cells<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e99"/> he described as "having the shape of a sickle." When Herrick saw this in the chart, he became interested because he saw that this might be a new, unknown <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e102" spanTo="#d0e102"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e102"/>. He subsequently published a paper in one of the medical journals in which he used the term "sickle shaped cells."

Originally from Africa and brought to the Americas by the forced immigration of slaves, it is more frequent where the proportion of African descendants is greater. Carriers of the <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e108" spanTo="#d0e108"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e108"/> trait have some resistance to the often-fatal malaria. This is why it is found more frequently in persons of Middle Eastern, Indian, Mediterranean and African heritage because those geographic regions are most prone to malaria.

However, in areas such as the U.S., where malaria is not a problem, the trait no longer provides a survival advantage. Instead, it poses a threat in children of carriers who inherit the <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e114" spanTo="#d0e114"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e114"/> gene from both parents.

There are several types of <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e120" spanTo="#d0e120"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e120"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e123" spanTo="#d0e123"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e123"/>, and the most common one is <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e126" spanTo="#d0e126"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e126"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e129" spanTo="#d0e129"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="sd0e1292" spanTo="#d0e1292"/>anemia<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="d0e1292"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="d0e129"/>. It makes you feel tired, because your blood cells don't last as long as other people's do.

Normally, <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e135" spanTo="#d0e135"/>red blood cells<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e135"/> live about 120

days. New ones that replace them are made in the soft, spongy center of your bones called the marrow. If you have <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e138" spanTo="#d0e138"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e138"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e141" spanTo="#d0e141"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="sd0e1412" spanTo="#d0e1412"/>anemia<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="d0e141"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="d0e1412"/>, your <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e144" spanTo="#d0e144"/>red blood cells<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e144"/> start dying after only 10 to 20 days. Your bone marrow can't replace them fast enough, which causes <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e147" spanTo="#d0e147"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="d0e1472" spanTo="#d0e1472"/>anemia<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="d0e1472"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="d0e147"/>. <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e150" spanTo="#d0e150"/>Red blood cells<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e150"/> carry oxygen around your body, which gives you energy. If you don't have enough of them, you'll get tired and you might also feel short of breath.

SCD <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e156" spanTo="#d0e156"/>patients<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e156"/> learn to function in a constant state of pain and when that pain becomes debilitating, they often end up in the emergency room. What brings people to the emergency room is called a crisis, in which the blood can't deliver oxygen to the extremities. It is a sudden onset of severe, excruciating pain, which people describe as feeling as if all their bones are breaking. The pain is so bad and sudden that people require high doses of pain killers.

Contrary to the belief of many in the medical profession, there is no data to support that <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e162" spanTo="#d0e162"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e162"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e165" spanTo="#d0e165"/>patients<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e165"/> have any more likelihood of being addicted than anybody else.

Acute and unpredictable pain, severe enough to require large doses of narcotics, is a well-documented feature of the <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e171" spanTo="#d0e171"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e171"/>. When many <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e174" spanTo="#d0e174"/>patients<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e174"/> arrive at emergency rooms, they often have great difficulty getting the <milestone

```
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e177"
spanTo="#d0e177"/>treatment<milestone xmlns="http://www.tei-c.org/ns/1.0"
n="r5" xml:id="d0e177"/> they need.</p>
            <p>Racism and the <milestone xmlns="http://www.tei-c.org/ns/1.0"
n="r5" xml:id="sd0e183" spanTo="#d0e183"/>disease<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e183"/> stigma itself are
two barriers that you just can't get away from. Clearly we can't pretend that
racism doesn't play some part in this. If this were a white <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e186"
spanTo="#d0e186"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="d0e186"/>, people still wouldn't be dying in their 40s. That's the
bottom line. <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="sd0e189" spanTo="#d0e189"/>Sickle cell<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e189"/> was "discovered"
106 years ago and there is only one drug, hydroxyurea, and blood transfusions
to treat it.</p>
            <p>There has been progress in <milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e195"
spanTo="#d0e195"/>sickle cell<milestone xmlns="http://www.tei-c.org/ns/1.0"
n="r5" xml:id="d0e195"/> <milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="sd0e198" spanTo="#d0e198"/>disease<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="d0e198"/>. People didn't
used to live to be adults. Kids would die of stroke or of really bad infections
before they were 20, and in some countries they still do.</p>
        </block>
    </body.content>
</Publication>
```

B.5 Hit Highlighting Example Discussion

Let's look at a couple of examples. These two `<milestone>` elements span the word "disease". And they reference the clause identified as r5 in the rule (via the `@n` attribute).

```
<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5" xml:id="sd0e198"
spanTo="#d0e198"/>disease<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r5"
xml:id="d0e198"/>
```

Here is an example where a single word ("anemia") is wrapped by four `<milestone>` elements. That's because it matches two different clauses in the rule: r1 and r2.

```
<milestone xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="sd0e147"
spanTo="#d0e147"/><milestone xmlns="http://www.tei-c.org/ns/1.0" n="r2"
xml:id="d0e1472" spanTo="#d0e1472"/>anemia<milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r2" xml:id="d0e1472"/><milestone
xmlns="http://www.tei-c.org/ns/1.0" n="r1" xml:id="d0e147"/>
```

DOCUMENT HISTORY

2017/01/30 - Version 1.0 - First non draft version.
2016/11/08 - Version 0.22 - Removed all references to miss highlighting and added stand-off markup as an alternative to milestones.
2016/11/03 - Version 0.21 - Added Appendix B to illustrate hit and miss highlighting.
2016/10/17 - Version 0.20 - Added a Table of Contents.
2016/10/06 - Version 0.19 - Rewrote the input requirements to better explain the “parameters” idea.
2016/10/03 - Version 0.18 - Added page numbers. Incorporated sample rules as Appendix A.
2016/09/27 - Version 0.17 - Clarified the differences between “default” and “custom” and “core” and “nice-to-have” relevance requirements.
2016/09/21 - Version 0.16 - Rewritten and expanded discussion of relevance algorithm
2016/09/12 - Version 0.15 - Moved hit/miss highlighting to be next to relevance, for clarity
2016/09/07 - Version 0.14 - Incorporated input and output format requirements.
2016/09/07 - Version 0.13 - Introduced hit and miss highlighting as their own APIs, distinct from classification.
2016/08/31 - Version 0.12 - Imported requirements from rule language requirements document.
2016/08/19 - Version 0.11 - Descoped automated conversion other rule languages; fixed numbering of use cases; require use of UTF-8, rather than UTF-16; tied up Nice to Have section.
2016/08/12 - Version 0.10 - Added more specifics to the “Performance” section; clarified Natural Language requirements; added schema, dictionary and relevance algorithm management use cases
2016/08/02 - Version 0.9 - Gave “relevance” its own section. Moved “hit highlighting” into “Interaction”. Added the specific list of natural languages to be supported. Added a “Machine Learning” section. Filled in “Background” and “Assumptions”.
2016/07/27 - Version 0.8 - Clarified relevance requirements; introduced “Richer UI” requirements section.
2016/07/19 - Version 0.7 - Moved rule testing with a scored corpus into “nice to have”.
2016/06/30 - Version 0.6 - Added new “nice to have” section.
2016/06/29 - Version 0.5 - Added new user stories - “hit highlighting” and “rule match scoping”. Introduced new sections to cover format requirements for the documents to be classified and the rules language.
2016/06/09 - Version 0.4 - Added one new user story - “calculate document relevance” and clarified that all features need to be documented.
2016/05/24 - Version 0.3 - Added ability to set and edit metadata per rule, but decided that filtering/searching is out of scope; clarified that rule testing returns precision and recall scores; first fill in of UI section; indicated that rule versioning is out of scope.
2016/05/13 - Version 0.2 - Removed rule set concept (instead, rules can reference other rules and you can request classification using a particular id). Clarified a lot of wording. Altered target year to 2017 from 2016!

2016/04/28 - Version: 0.1 - First version.