

R3 Release Notes

What's in R3

Calais Release 3 (R3) includes new Entities, new Events/Facts, additional attributes for existing Events/Facts, numerous new features, new interfaces (REST), and enhancements to Gnosis, Simple Formats, and Microformats.

R3 Features - Highlights

| | |
|--|---|
| New Entities | See table below. |
| New Events/Facts | See table below. |
| New Attributes for existing Events/Facts | See table below. |
| Generic Relations | <p>This module provides exhaustive extraction capabilities for all relations that involve at least one recognized entity (for example: Person, Company, TVShow, MedicalCondition etc., etc.).</p> <p>While the other events and facts extracted by Calais are predefined and usually relate to a specific domain (e.g., Mergers and Acquisition for the Business domain), Generic Relations attempt to recognize all the Subject-Predicate-Object relations without predefining their type.</p> <p>The Generic Relations module can be enabled by submitting the following parameter (processing directive) in the API call:</p> <p style="text-align: center;"><code>enableMetadataType="GenericRelations"</code></p> <p>To get GenericRelation results you must choose the RDF output option; results are not included in other output formats.</p> |
| Categorization | <p>This new capability support document categorization – the ability to recognize the topics or “aboutness” of submitted content. In this version we deployed a general taxonomy of six topics, which will be expanded in future releases.</p> <p>The topics supported in this version are: Business/Finance, Sports, Entertainment, Health/Medical, Politics and Technology.</p> <p>Categorization results appear both in RDF output and Simple Format output.</p> |
| Relevance | <p>This module includes several enhancements:</p> <ul style="list-style-type: none"> ▪ Entity scores now have 3-digit precision after the decimal point ▪ Entity scores are now comparable across input texts. This means that you can use entity relevance scores in order to determine entity relevance at a collection level – not just document level. (In previous releases scores were normalized at the document level, hence not allowing the comparison of entities across multiple texts) |
| Gnosis | Ability to sort entities by Relevance (Firefox only) |
| Microformats | <p>Enhancement enables insertion of Microformats results directly into an HTML page without affecting page layout – that is, adding the microformats for each entity so that entities/events can be visually highlighted by Microformats-aware browsers (plugins to Firefox).</p> <p>This was done by using the tag in the Microformats output.</p> |

| | |
|-------------------------------|---|
| Interfaces | <p>REST interface to Open Calais for those using PHP/Java/JSON is now supported. The following is a sample REST API call via HTTP POST request and response. The placeholders shown need to be replaced with actual values. Please note that the content sent using this method needs to be escaped.</p> <p>The request itself is identical to HTTP POST, but it should be sent to a different URL. The response does not include an enveloping XML with a <string> element, it is the RDF itself.</p> <pre>POST /enlighten/rest/ HTTP/1.1 Host: api.opencalais.com Content-Type: application/x-www-form-urlencoded Content-Length: length licenseID=string&content=string&paramsXML=string HTTP/1.1 200 OK Content-Type: text/xml; charset=utf-8 Content-Length: length string</pre> |
| calaisRequestID | Now available (was previously announced as CalaisSessionID) |
| Timeout and partial results | <p>A timeout of 20 sec is applied if large input content is submitted to Calais. However instead of dropping the transaction, Calais will return the results extracted so far in the RDF output, and will also indicate the occurrence of a timeout for the submitted content.</p> <p>The message in the RDF output will be as follows:</p> <pre><c:message> <rdf:Description> <rdf:type rdf:resource="http://s.opencalais.com/1/type/sys/Message" /> <c:messageCode>201</c:messageCode> <c:text>Partial metadata extraction due to timeout </c:text> </rdf:Description> </c:message></pre> |
| Secured Connection | Calais now supports SSL security of traffic to and from Calais. GoDaddy is the authority for SSL certification. |
| Input Content Language | <p>Calais today supports English only. To ensure non-English content is not processed, Calais applies a Language Identification module before processing the text for entities, events etc.</p> <p>So far if the text was too short the language wasn't recognized and hence no extractions were returned.</p> <p>In this version, if the submitted text is too short (less than 100 characters) and the language can't be recognized, Calais will assume the language is English by. In addition, in such cases, Calais will return "InputTextTooShort" as the language code in the RDF.</p> |
| Efficiencies and improvements | Lower latency in Simple Output format and Microformats |

New Entities

| | |
|---------------------|---|
| MedicalTreatment | <p>Extracts references to medical treatments - procedures, treatments and therapeutics provided to any medical condition.</p> <p>For example: Merck plans to conduct a Phase 3 trial of oral deforolimus in patients with metastatic soft-tissue and bone sarcomas following a favorable response to chemotherapy.</p> |
| OperatingSystem | <p>Extracts references to operating systems.</p> <p>For example: Jay reported yesterday that the Google Maps for mobile for Palm OS update would be available today</p> |
| ProgrammingLanguage | <p>Programming Language</p> <p>Extracts references to programming languages.</p> <p>For example: Resolver One uses the Python programming language, which was first released in 1991.</p> |
| SportsLeague | <p>Extracts names of sports leagues. Note that some sport leagues may also be extracted as organizations.</p> <p>For example: Taylor Smith discovers that there are Major League Baseball teams in Florida</p> |

New Events/Facts

| | |
|-------------------|--|
| BonusShares | <p>Extracts references to bonus shares issues (in the past, present or future).</p> <p>For example: Malaysia's Public Bank said on Monday it plans to issue one bonus share for every four existing.</p> <p>The extracted attributes are: Company, Ratio, Date, DateString, Date_Issued and Status</p> |
| CompanyTicker | <p>Extracts instances where the company name is followed by its stock symbol/s (ticker symbol/s) in parenthesis.</p> <p>For example: Baltimore Technologies (NASDAQ:BALT; London:BLM)</p> <p>The extracted attributes are: Company, CompanyTicker and StockExchange.</p> |
| MovieRelease | <p>Extracts references to movie releases (past, present or future) or mentions of new movies.</p> <p>For example: The latest Will Smith movie, 'Hancock,' will hit theaters Wednesday, July 2.</p> <p>The extracted attributes are: Movie, Date, DateString and Status</p> |
| MusicAlbumRelease | <p>Extracts references to music album releases (past, present or future) or mentions of new music albums.</p> <p>For example: The Mission released its latest album, "This Violent Decline", in the U.S. on August 28, 2007 via Locomotive/Dockyard 1/Ryko Distribution</p> <p>The extracted attributes are: MusicAlbum, Person_Performer, MusicGroup_Performer, Date, DateString and Status</p> |

New Attributes for Existing Events/Facts

| | |
|-----------------------------|--|
| AnalystEarningsEstimate | FinancialMetric - new attribute |
| CompanyEarningsAnnouncement | FinancialMetric - new attribute |
| CompanyEarningsGuidance | FinancialMetric - new attribute |
| AnalystRecommendation | Price_New and Price_Old – new attributes |
| BusinessRelation | BusinessRelationType – new attribute |
| StockSplit | StockSplitRatio – new attribute |
| CompanyAffiliates | AffiliateRelation – new attribute |
| CompanyLocation | LocationType – new attribute |

Known Issues

Submitted Content – Categorization will produce erroneous results if the submitted text is too short. Please make sure the submitted content has at least 1K of text.

Secured Connections via HTTPS – To use secured connections via HTTPS, the Java installation has to be JRE 1.4.2_07 or higher and 1.5.0_02 or higher. Older versions will most likely produce the exception `SunCertPathBuilderException`. If you can't upgrade your Java installation, you can manually import the Calais SSL certificate to your key store (consult the keytool documentation for your specific Java version).

To bypass the certificate authority in PHP, in case it is not supported by the software package, do the following:

- **JSON Simple Format:** Set `$simplejson_VerifySSLCertificates=false`; in `json\simple\conf.php`
- **JSON Full Format:** Set `$fulljson_VerifySSLCertificates = false`; in `json\full\conf.php`
- **Marmoset:** Set `$calaismf_VerifySSLCertificates = false`; in `calaismf\conf.php`

TEXT/RAW Input – This is the default in Calais, and can be overridden by the appropriate processing directive.

When the input format is TEXT/RAW, the offset/length is not correct when referring to the input text as it appears in XML/RDF CDATA (the text in CDATA is escaped). The offset/length is correct when referred to in the original document sent to Open Calais.

The format of `paramsXML` needs to be a full XML and not a short one. Use:

```
<c:processingDirectives .... > </c:processingDirectives>
```

instead of:

```
<c:processingDirectives .... />
```

This issue will be addressed in the next release.

Other Resources

One of the first places to look for additional information is the Forums on the Calais website (www.opencalais.com/Forums), where Calais users share their knowledge and expertise, and questions are answered by the Calais team.