|  |  |  |  |
| --- | --- | --- | --- |
| Item | Name | Component Description | |
| 0 | OpenSextant Commons | Often used data model for abstracting gazetteer data and the core attributes about textual annotations; As well as any common text and IO utilities used across 2 or more OpenSextant modules | |
| 1 | Gazetteer Processing | This component is responsible for the fetching, cleaning, and transformation of the source gazetteer data sets into a single clean consistent gazetteer data set ready to be used by the geotagger | |
| 1a | Gazetteer data | This is the actual gazetteer data produced by the gazetteer cleaning/transforming process. This may take the form of either a CSV file or a prebuilt Solr gazetteer (see item #2) | |
| 2 | Gazetteer Repository | This component stores the gazetteer information. Built on Solr. | |
| 3 | Solr Matcher | This component provides the document-to-gazetteer matching for the Solr repository. It is identified as a separate component since it may be contributed to the Solr project baseline (independent of any OpenSextant open sourcing effort) | |
| 4 | Xponents | These set of components provide the implementations for the finding geocoords (Xcoords) and date/time information (Xtemporal). Further components are planned (details TBD). These are identified as separate components since they can be used independently as standalone coordinate and date/time extractors for use cases which do not need or cannot support the full OpenSextant extraction processes. | |
| 4a | FlexPat | A grammar based on REGEX for defining and testing pattern-based extraction which yields higher level text annotations | |
| 4b | Xtext | A component used for acquiring text from data artifacts; Captures important metadata, text payload properly in UTF-8, and other mechanics of input files; Xtext offers a way of persisting organizing conversions on disk as well | |
| 4c | Xcoord | Geographic coordinate extraction and normalization | |
| 4d | Xtemp | Date and time extraction and normalization | |
| 5 | OpenSextant Toolbox | This component consists of 8 Natural Language Processing (NLP) components (called ProcessingResources) which can be combined with other ProcessingResources provided by the GATE framework (see GATE dependency below) to build a complete NLP processing pipeline (such as items #7 and #8 below). These ProcessingResources are described on the OpenSextant Toolbox Components table below. | |
| 6 | Language Resources | This component consists of the rules, patterns, vocabularies and other language data resources used by the OpenSextant Toolbox (this components is data only, no software) | |
| 7 | Geotagger Pipeline | This component is the definition of the complete sequence of steps needed to perform geotagging. It consists of a GATE application definition file (XML) which specifies which ProcessingResources (see #5 OpenSextant Toolbox above) and Language Resources (see #6 Language Resources above) are to be used in what order to perform geotagging. | |
| 8 | General Purpose Extractor Pipeline | THIS COMPONENET IS UNDER DEVELOPMENT. This component is the definition of the complete sequence of steps needed to perform general purpose entity extraction It consists of a GATE application definition file (XML) which specifies which ProcessingResources (see #5 OpenSextant Toolbox above) and Language Resources (see #6 Language Resources above) are to be used in what order to perform general purpose extraction | |
| 10 | Applications | It is expected that there will be numerous applications which build open one or more OpenSextant components. Those built by or known to the OpenSextant project are listed below. | |
| 10a | OpenSextantRunner | A complete command-line based document processing/geocoding pipeline; Documents are ingested/converted/saved, gecoded… this would work fine as a batch processing for small ~1000s of documents | |
| 10c | SimpleGeocoder | A core geocoding class that provides access to the bare minimum, full geocoding pipeline | |
| 10d | OpenSextant Desktop | A desktop GUI for an end user to manage their own geocoding runs | |
| 10e | TweetGeocoder test | A demonstration of how to use a semi-structured data set, SimpleGeocoder, and the OpenSextant output module to generate GIS products live. | |
| 9 | Web Services | There is some value in figuring out a real full-up WS or REST interface and support it… but this will lilkely be the first thing the OSS community will take on… and others out there will do this much better than we. | |
|  | General Purpose Extractor webservice | THIS COMPONENT HAS NOT YEST BEEN BUILT. When built, this component will expose the general purpose extraction pipeline as a web service. Further details TBD. | |
|  | OpenSextant Geotagger web service | This component exposes the Geotagger pipeline as a web service. It also provided the capability to provide the results in a variety of geospatial output formats. This component also provides a basic user web application (user interface) to allow an individual to submit documents to be geotagged. | |
| Toolbox Component | | | Description |
| NaiveTaggerSolr | | | Matches gazetteer entries to document contents. Based on Solr. The primary NaiveTagger |
| PhoneticTagger | | | Tags tokens with phonetic or simplified version of the token. |
| Cantilever | | | A geospatial specific form of coreferencer. |
| CaseDetector | | | Determines if a document is written in all upper, all lower or "proper" case |
| GeocoordFinder | | | Tags gecoords in documents. Built on top of Xcoord |
| DateFinder | | | Tags date/time information in documents. Built on top of Xtemporal |
| POSTagger | | | IN PROGRESS,REWORK FROM EARLIER FORM TO OPENSEXTANT FORM. Tags tokens with their part of speech. Used by the General Purpose extraction process |
| Decapitator | | | IN PROGRESS,REWORK FROM EARLIER FORM TO OPENSEXTANT FORM . Converts documents written in all upper, all lower or inconsistent case to "proper" case. |

Toolbox components

|  |  |  |
| --- | --- | --- |
|  | GATE | External Open source dependency (GATE Project) GATE is an open source NLP framework which provide the basic infrastructure for document processing, text analysis and entity extraction. The OpenSextant Toolbox (item #5 above) extends this framework by providing ProcessingResources needed for geotagging and general purpose extraction |
|  | Kettle | External open source dependency (Pentaho Inc) Kettle (aka Pentaho Data Integration) is an Open Source Extract-Transform-Load application which the OpenSextant project uses to process, clean and transform the gazetteer data. |
|  | giscore | External open source dependency (MITRE giscore project) giscore are a set of Java libraries which can read and write geospatial data formats such as SHAPE files, KML and ESRI File Geodatabases. These libraries are used by OpenSextant to create the results returned by the geotagging process. |

Major External dependencies