|  |  |  |  |
| --- | --- | --- | --- |
| **Open Source Knowledge Enrichment (OSKE)**  Open Source (OS) Knowledge Enrichment (OSKE) is a system to manage the complexities of capturing, formatting, manipulating, and understanding publicly available information. The internet-connected platform hosts techniques and analytics designed to be semi-automated and allow for the configurability of querying a wide range of structured and unstructured data sources. The system helps capture the ”right” (relevant) information from the ”right” (reliable) sources, and the system provides capabilities to holistically analyze big data within an internet-connected environment and integrate data of value into the mission environment. *Note: OSKE is also referred to as OpenKE.*  OSKE is organized into *Domains*, which serve as an overall information divider between specific research topics. Within each *Domain*, users can focus their interests into a *Project.* For example, a “cybersecurity” *Domain* could exist with two *Projects*: “electrical grid security” and “IOT security”.  Within a *Project*, users can also create and maintain *Scratchpads* in order to document important items they have discovered as well as to take notes.  **Workflow**   |  |  | | --- | --- | |  | OSKE’s workflow design (Figure 1) is based upon a common, iterative analyst workflow; it includes the following steps:  **Plan**: Enables users to outline and document intended research objectives, assumptions, and key questions.  **Discover:** Enables users to initiate searches to query publicly available information (PAI) on subject/topic area of interest.  **Retrieve:** Enables users to create and execute iterative and standing recurring queries to retrieve PAI.  **Analyze:** Enables users to sort, select, visualize, and begin to make analytic judgements on query results.  **Document:** Enables users to annotate queries and create internal notes, capturing interesting findings.  **Manage:** Enables users to adjust retrieval strategy to tailor retrieval and bring in new PAI over time. |   *Figure 1: OSKE Workflow* | |
| **OSKE Home Page**  The OSKE Home Page (Figure 2) lists the *Domains* available to a user within the system and displays recent information retrievals. | |
| **Area 1** contains the hyperlinked Name for each available *Domain*. Users start here by selecting a Name from the listing to begin their research and navigate to that particular *Domain*’s dashboard.  **Area 2** contains a description of each domain.  **Area 3** lists the point-of-contact for the domain.  **Area 4** shows a graphical representation of recent retrieval jobs within the last 12 hours.  **Area 5** is the OSKE icon, which is a hyperlink on all OSKE page’s to return the user back to the OSKE home Page (Figure 2). | *Figure 2: OSKE Home Page* |

**Domain Dashboard Page**

|  |  |
| --- | --- |
| The Domain Dashboard Page (Figure 3) provides a high-level overview of the currently-selected *Domain*. | |
| **Area A** is the navigation bar used to access the workflow tabs. Selecting Dashboard tab from this area will direct the user back to the Domain Dashboard Page.  **Area B** details the current workspace. The current *Project* name and *Scratchpad* name are listed.  **Area C** includes the Home View, Logout icon, and Feedback form. The Home View allows the user to toggle between the Analyst view (default) and Collector view, which makes examining jobs and retrieved pages easier. The Logout icon (LAS) will exit OSKE. The Feedback form is used to submit bug reports or request support.  **Area D** contains the system version (build) and the current time.  **Area E** lists the *Projects* in the current *Domain*.  **Area F** summarizes existing sessions within the current *Domain*. If there are no sessions, the table will have no entries. Users start a new session here by entering search terms, selecting a source handler, and clicking the *New Session* button. This will direct the user to the *Discover* tab where the results will be displayed.  **Area G** displays recent news from a default set of news sources. News sources can be customized by selecting the gear icon. | *Figure 3: Domain Dashboard Page*  **Area H** is the Search Alerts pane. This pane displays any new results from recurring queries. To clear a single alert, click the *X* icon next to the alert. To clear all alerts, select *Acknowledge All*.  **Area I** presents concept frequencies from collected information.  **Area J** links to reference materials for OSKE. |
| Helpful Hints:   * Typically, the OSKE workflow begins with *Plan*. Select the *Plan* tab from within Figure 3 Area A. If a *Plan* has not yet been created, the Project Plan form must be completed. Refer to subsequent Project Plan Page for more details. * Select the *Discover* tab from within Figure 3 Area A. To view a past session, select a Name from the Session table. If a session has not yet been created, select the *Start New Session* button. Refer to subsequent Domain Discovery Session Page for more details. * Search alerts may be created to identify whenever a new result from a search engine appears for a given set of terms. * To continually re-run a query to identify new results of interest, a Search Alert may be created. To create a new Search Alert, the Create Search Alert form must be completed. Select the drop-down icon of the *Manage* tab from within Figure 3 Area A; then select *Search Alerts*.   + Select the *Create Search Alert* button. The Create Search Alert dialog box will appear with the following fields:     - **Search Alert Name:** Enter a meaningful name for the alert.     - **Search Terms:** Enter any keywords or keyphrases to be searched.     - **Number of Search Results:** Select the number of results to be retrieved. The default is 20.     - **Schedule:** Select the frequency to conduct new searches. The default is every 8 hours.     - **Method:** Select a source handler to use for the search.     - **Acknowledge Initial Results:** If desired, select the box to clear alerts from the Dashboard view. The default is unchecked.   + Once these fields are complete, select the *Create* button. | |

**Domain Discovery Session Page**

|  |  |
| --- | --- |
| The Domain Discovery Session page (Figure 4) allows the analyst to perform a variety of different searches, view in-line summaries of documents, look for new search terms by clustering documents in *Topic Modeling*, and navigating through results based upon keywords (i.e., a “*book*” index) or identified concepts. | |
| To begin a new session, the user must complete a Discovery Session Query form (Figure 4 Pane A). There are four sections of the form:  **Session Name:** Input a meaningful name for the session.  **Search Terms:** List search terms to be queried.  **Number of Search Results:** Select the number of results to be retrieved. The default is 20.  **Method:** Select the Source Handler to be used.  Once the form is completed, select the *Search* button to submit. This populates the *Search Results* pane (Figure 4 Pane B). By selecting the text for a specific result (or the *plus* icon), a dynamic text summary will appear. Users can adjust the length of the summary with the slider (Figure 5). Users can toggle between List View (shown in Figure 4 Pane B) or Index View (shown in Figure 5) by selecting the Switch to Index View or Switch to List View (location highlighted in Figure 5 Area 1). The *Topic Modeling* section will populate with clusters of documents. Click a topic keyword to add to the search terms.  Note-taking is provided by the *Copy-Paste-to-Scratchpad* feature. Within the Results Pane, highlight the text to be saved to the *Scratchpad*. The Create Note dialog box will appear with the Selected Text field populated with selection. Add any notes to the Your Note field. Select the *Confirm* button. This will send the text and note to the current *Scratchpad.* Refer to *Scratchpad* page for more details.    To create a scheduled job, click the *Create* button and select Job. After entering a justification, the user will be directed to the Add Job form (refer to Add/Edit Job Page for more details). | *Figure 4: Domain Discovery Session Page*  *Figure 5: Domain Discovery Search Results Page* |

**Project Plan Page**

|  |  |
| --- | --- |
| As the name implies, the purpose of the *Plan* step is to provide a method of planning the user’s intended research concepts and goals. | |
| Figure 6 Pane 1 is the Projects Pane, and it displays the available *Projects* within the *Domain*. If there are none available and/or selected, it may read “No active projects” or “No project is currently select or set active”.  To begin a new *Project*, the user must fill out the Project Detail form (Pane 2). Select the *Create* button at the top-right of Pane 1. There are five sections of the form:  **Name:** Input a meaningful name for the research project.  **Purpose:** Provide a description of the project.  **Key Questions:** State the main questions that the project aims to answer.  **Assumptions:** List all pre-existing beliefs or caveats.  **External Links:** Provide any URLs that may be relevant.  Once these five sections are complete, select the *Save* button at the top-right of Pane 2. | *Figure 6: Project Plan Page* |
| **Scratchpad Page**  The *Scratchpad* is a text editor, allowing the user to take notes on OSKE findings. | |
| Figure 7 Pane A is the *Scratchpad* pane, and it displays the available *Scratchpads* within a *Domain*. If there are none available and/or select, it may read “No active scratchpads” or “No document is currently selected or set as active”.  To create a new *Scratchpad*, the user must fill out the scratchpad details in Pane B. Select the *Create* button at the top-right of Pane A. In Pane B, provide a Scratchpad Name for the *Scratchpad* and enter any notes within the text editor (standard text features supported). Select the *Save* button at the top-right of Pane B. *Scratchpads* can be converted to Microsoft Word documents (via *Export*).  To mark a *Scratchpad* as the “current” *Scratchpad*, select a Scratchpad Name from within Pane A. Within Pane B, select the *Make Current* button at the top-right. | *Figure 7: Scratchpad Page* |

**Add / Edit Job Page**

|  |  |
| --- | --- |
| The *Add / Edit Job Page* (Figure 8) allows the user to configure a specific retrieval job that runs on a scheduled basis to retrieve PAI on a subject/topic of interest from the Internet. Users can use search engines, RSS feeds, URLs, and other items as the initial points to retrieve data. | |
| To modify an existing job, the Edit Jobs form must be completed. Select *Edit* of the specific job’s cell within the Jobs table. There are eight sections of the form:  **Name:** Input a meaningful name for the retrieval job.  **Search Terms/URL:** Enter the search terms used for to initiate the job. Alternatively, enter the starting URL, if using the RSS Feed Handler or Web Crawler.  **Source Handler:** Select the desired source handler used for crawling.  **Priority:** Enter a priority value from 1 (lowest priority) to 100 (highest priority). The default is 100.  **Schedule:** Decide when the retrieval job should run by using the drop-down selections.  **Randomize:** If necessary, select the Randomize box to vary the start time of the job and enter a percentage.  **Configuration:** Select the *Load Template* button to set the configuration to the default for the particular source handler. Otherwise, modify parameters as desired.  **Justification:** Enter the rationale for why this job is needed.  Once these eight sections are completed, select the *Submit* button at the bottom of the form.  Once a job has been submitted (including any edits), it must be approved by an adjudicator before it will execute. The adjudicator reviews jobs for policy and compliance. All jobs must adhere to policy guidelines as follows: 1) Abide by the website's rules for retrieving and using the data from that website. 2) Data must be publicly available with no special account access or log-on password required. 3) Users should make the best effort to limit and/or remove data containing any personal information, porn or illegal activity. | *Figure 8: Edit Job Form*  Some common settings of Configuration are described below:  **allowSingleHopFromReferrer:** Allows the crawl job to stay on a particular domain, but then bring in a page if it is referenced. Useful for retrieving source documents referenced on wikipedia pages.  **limitToDomain:** If true, the crawler will stay on a starting domain.  **relevantRegExp:** If a regular expression is used, crawling down a particular path will stop if this expression is not found in a retrieved document.  **webcrawler.maxPagesToFetch:** Maximum number of pages to retrieve for a job’s execution.  **webcrawler.politenessDelay:** Number of milliseconds before starting a new request.  **webcrawler.respectRobotsTxt:** Should be set to true. Ensure OSKE will follow crawling rules identified in a site’s robots.txt file |

**Analyze Tab**

|  |  |
| --- | --- |
| The purpose of the *Analyze* step (Figure 9) is to provide a visual summarization of query results in order for trends and patterns to become apparent. | |
| By default, the *Analyze Tab* visualizes all retrieved data within a *Domain*. Users can apply filters to restrict the data analyzed.  The Frequencies visualization (Figure 10a) provides a view based on the number of occurrences of noted terms. The Y-axis are the most common terms, and the X-axis is the count of each term. By default, it will display the top 20 result terms.    The Geospatial visualization (Figure 10b) provides a view based on the geographic locations referenced in the text from the queried results.  The Heatmap visualization (Figure10c) provides a view based on custom X- and Y-axis parameters. This allows users to discover trends between seemingly-unrelated pieces of data. The Heatmap displays the number of co-occurrences between the defined parameters.  To generate a Heatmap, select the desired X- and Y-axes from the drop-down menus (Figure 10c Area A). Then, select the *View* button.  The Filter Dialog (Figure 11), accessed by selecting the *Filter* button at the top-right of Figure 9, allows the user to limit the data used within the visualizations. Select terms (i.e., entities) to filter upon by clicking on them. Use *shift-* and *command*- click gestures to select multiple items. Select *Update by Filters* to update the current categories based the data and search terms selected. This technique provides an analytic in and of itself. Select *Apply Filter* to use the filters on visualizations. Select *Search By Filters* to see only documents matching the filter criteria.  To add categories, select an entity and/or concept option from the drop-down menu (Figure 11 Area 1). To remove a category, click the *X* at the top-right of the table. Once the desired tables are generated, click the *Update Filters* button. | *Figure 9: Analyze Tab HomePage*    *Figure 10a: Frequencies Figure 10b: Geospatial*  *Figure 10c: Heatmap*    Figure 11: Filters Selection |