Search requirements and roadmap

This page highlights the key features of Search, to help communicate, clarify and formalize the Vision for Search, so that the right technology choices are made by the AI and Engineering teams. The initial recommendation was Microsoft Azure Search with the Text Analytics Cognitive Search, but Nigel has strongly recommended ElasticSearch with Kibana.

Use Cases:

1. **Basic Search:** A user should be able to search across all document types and find information:
    a. Document types should include (digital and scanned PDFs, Microsoft Office suite (Word, PowerPoint, Excel, Outlook)
    b. The results should show words highlighted with page number of the document, with the link to go back to the source document
    c. The user should be able to export the results in a CVS or Excel format
    d. The document source for ingestion can be file uploads, file systems, blob storage in Azure, databases (Mongo, SQL, Cosmos), SharePoint libraries etc.
2. **Faceted Search:** A user should be able to have facets available to them from the basic search, which
    a. Shows them entities (People, Organization, Places, Document types, keywords etc.) that are displayed as facets with data filling in dynamically
    b. The users should be able to see how many documents/hits across each element of the facet, and be able to narrow down the results from the facets
    c. Allow the user to be prompted with words and phrases that are in the document or from earlier searches once they start typing (Type-ahead Search)
3. **Search with a dashboard view:** A user should be able to search across documents, by potentially an Entity (e.g. Adobe) and get a dashboard view of things like:
    a. Types of documents - Contracts, Invoices, MSAs, Amendments, etc. and total number of documents
    b. Drill down the invoices or contracts, and see if we can get any aggregation on Total value of contracts or number of licenses etc.
    c. Get licenses that are up for renewal within the next 3 months, 6 months etc. - a time series display with option to filter by date
    d. Identify contracts by region or country and show a map view of the same, to be able to drill into.
4. **Natural Language Query Search:** A user should be able to query using NLQ, and see results in a basic search results tabular format, or dashboard like format, based upon the query
    a. e.g. Show me the contracts with $1M or more as value and that are up for renewal within 3 months, by region or state
5. **Custom dashboards for different engagements/domains:** As each engagement that "Document Intelligence: Search" is used will vary on the domain and sector, the dashboard elements cannot be static.
    a. An Admin user at the Engagement should be able to define what are the types of entities they are going to be interested in (can pick from a list), and define any custom entities for facets (e.g. Business Unit)
    b. An Admin user should be able to set up dashboard elements for their dashboard - for eg. A pie Chart for document types, a bar chart for regions or state with contracts relevant to that region or state

Note: I have tried to abstract out the requirements to be somewhat tool agnostic, and have produced this table, which can help facilitate and finalize Product vision, priority and finding the right Engineering resources.

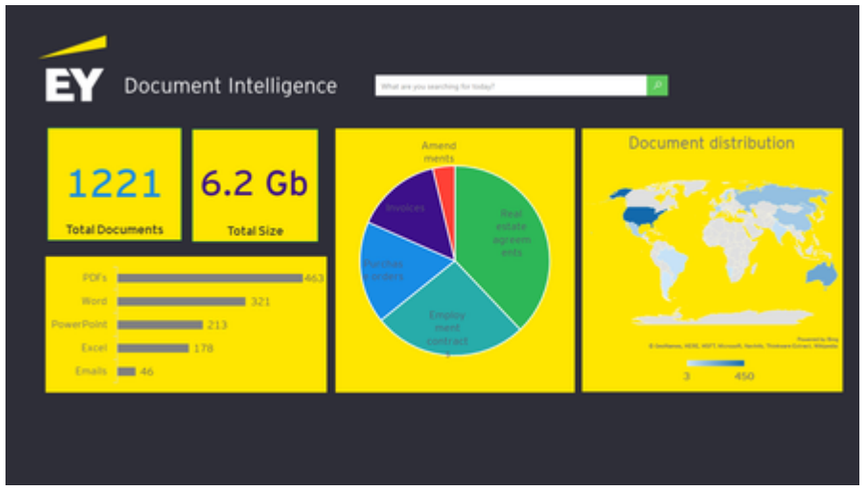
Search Features	Current DDP	Search MVP	Search V2	Search V3
Document Types	PDFs only	PDF, MS Office – Word, PPT, Excel	Outlook	
Document sources	File Uploads	File uploads	SharePoint library, External web repos, File Shares etc.	Databases, CMS, websites etc.
Full text search with Boolean operators and lemmatization and phrases	Yes	Yes		
Hit highlighting and page number of content	Yes	Yes		
Preview partial page/context and full document	Yes	Yes		
Suggester/Typeahead	No	Yes		
Semantic Search – Synonyms, Hyponyms and aliases with Ontology	No		Yes	
More like this (Term vectors)	No		Yes	
Interactive filters/Faceted search E.g. Person, organization, place, date, document type etc.	No	Yes		
Aggregation and Visualization E.g. Document dates as trend charts or text analysis as word clouds, connections and networks in visual graph view or view results with geodata as interactive maps.	No	No	Yes	Yes (may not have maps)
Collaborative annotation & Tagging	No	No	Yes	
Natural Language Query	No	No	?	Yes
To be able to filter or export	Yes	Yes		

Example Dashboards

Document Intelligence mockups

Key features:

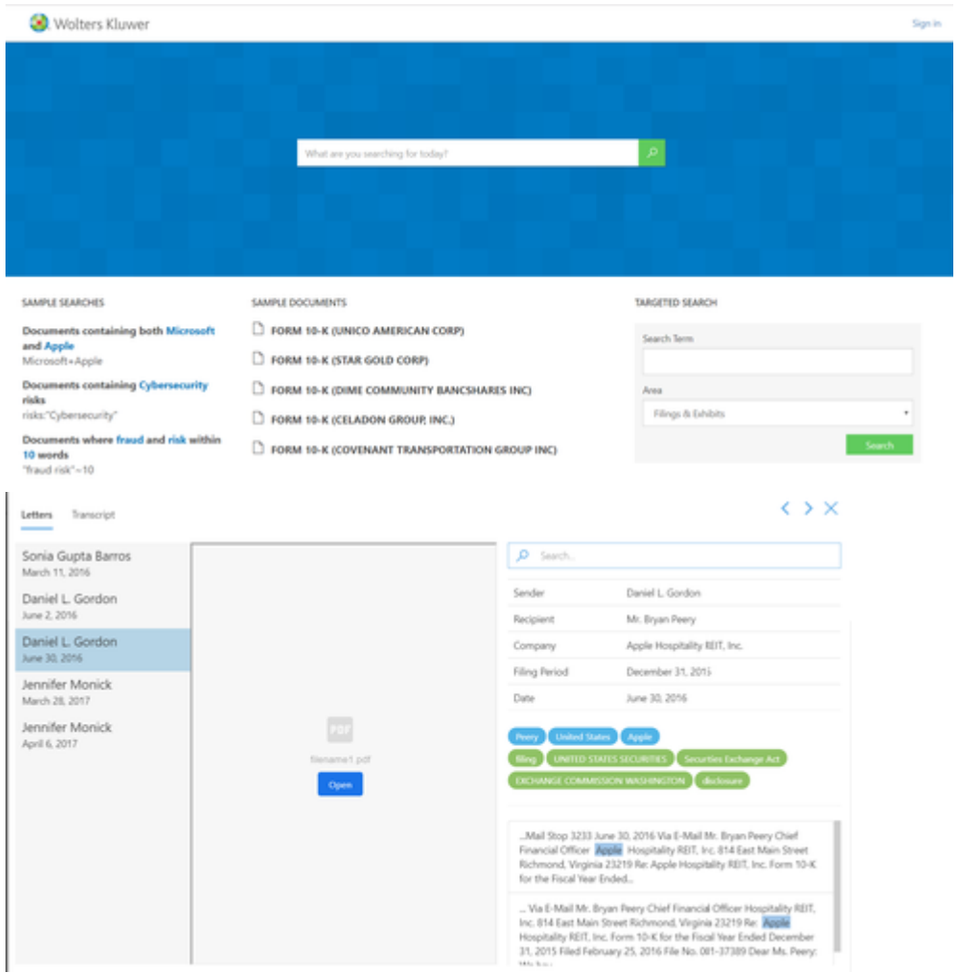
- Total number of documents, size, type etc.
- Classification facet and total docs
- Geo distribution from possibly the meta data of the document, stored in a database

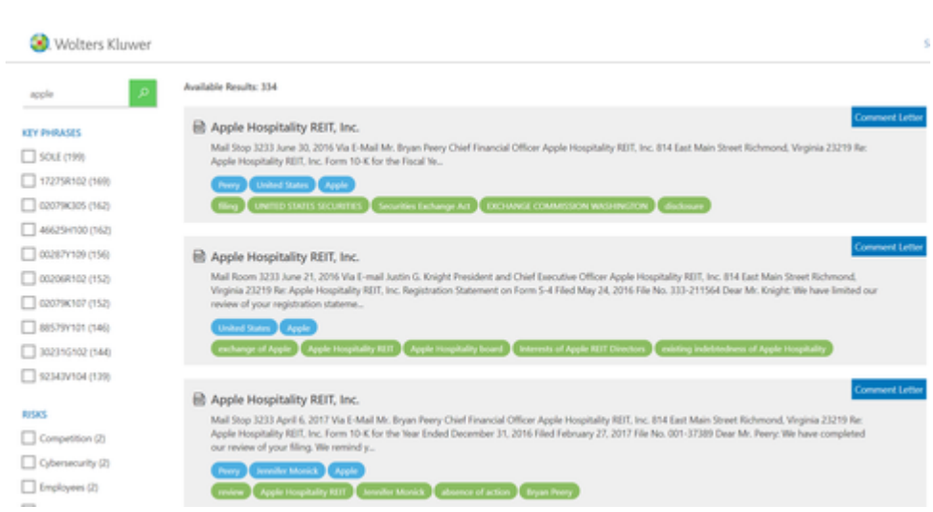


Microsoft Azure Search demo

Key features:

- Search (with suggestions) and an area.facet drop down, possibly from a Classification or document type
- Search results with highlights within the document, actual PDF or image, location of search term, some meta data of the document
- Facets for search, paragraphs or snippets of content with search term highlighted

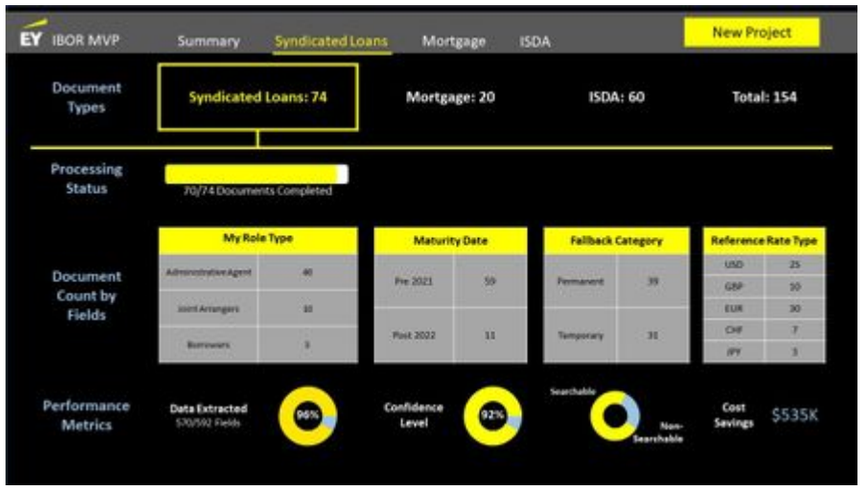


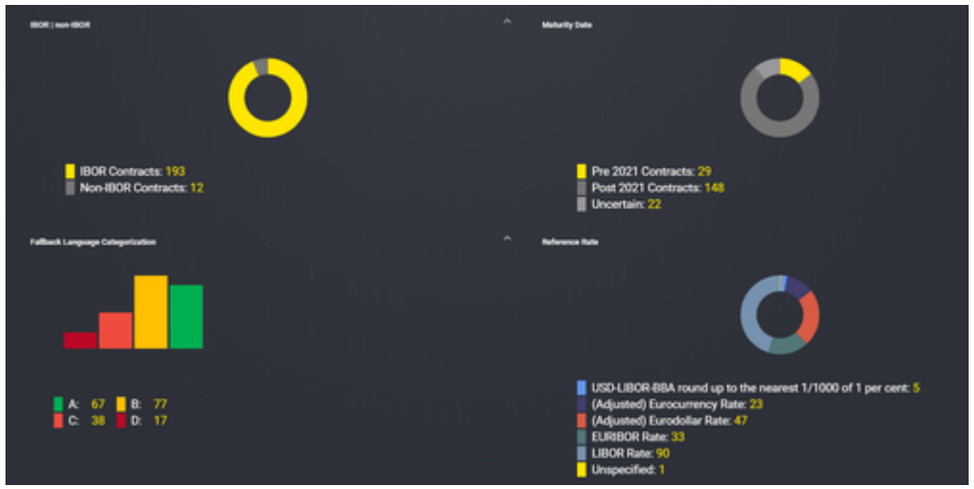


IBOR Dashboards

Key features:

- Total number of documents by type
- Extraction fields from a meta data table and corresponding counts
- Donut charts and other graph elements with metrics, also from the meta data table





LucidWorks Demo app screens

Key features:

- Standard Search elements with facets, type etc.
- Right nav bar with external content and meta data
- Custom tagging and bookmarking of documents, Saved Search terms
- Preview of document with Author and other relevant documents by that Author
- Geo mapping of people/authors and other facets of Search