Glossary of Terms Related to Search and Text Retrieval

Term	Definition	Synonyms
Aggregation	Activity for forming distinct sets of content	
Analytics	Data that helps track business trends Records that describe part of a larger domain Sophisticated version of data	Text analytics
Application programming interface	Vendor supplied add-on software tools to facilitate programming new features or functional enhancements to integrate a software product with other applications	API
Authority	Validating entity	
Authority control	Methods and lists employed for validating terminology and other content normalizing values in data maintenance	
Auto-categorization		See Categorization
Boolean searching	Use of explicit commands to limit or narrow the scope of a search (AND), expand its scope (OR), or exclude explicit content (NOT). e.g. search for content limited to containing both "energy" AND "solar", where AND is the command.	
Business analytics	Technique to visualize and analyze business data to support decision making	
Business intelligence	Technologies that gather, store, analyze and make accessible data to help enterprise users make better business decisions. It includes decision support, query and reporting, online analytical processing, statistical analysis, forecasting and data mining. Enhancing data into information and then into knowledge. Traditionally focused on extracting and manipulating data from structured databases including numeric data. Viewed by some as the umbrella for other technologies including text mining and analytics.	BI
Categorization	A computational or human activity assigning labels to sets of content to explicitly aggregate by label	
Citation	Information that accurately defines and describes a publication or data file;	SEE ALSO Results

	structured bibliographic metadata	
Clustering	Process for gathering unstructured content into a common space for the purpose of grouping it with content on the same topic	SEE ALSO: Aggregation Categorization
Collaboration	Describing shareable processes and/or content within an application.	SEE ALSO: Social search
Concept search	Retrieval of content through automated means that take contextual information, not just key words, into account when determining the relevancy of the content.	SEE ALSO: Semantic search
Connectors	Software tools supplied by search vendors or built internally to support data exploitation by a search engine.	Adapters
Content	The target of search regardless of format or medium. Everything included in a collection of files	
Context	Surrounding content that elucidates and clarified a set of data	
Controlled vocabulary	Terminology from approved lists used for tagging content	SEE ALSO: Taxonomy| Thesaurus|Ontology
Crawling/Spidering	Computer programs, usually part of a search engine, that traverse a specified set of domains for the purpose of indexing all content encountered	SEE ALSO: Indexing (Computer)
Cross-reference	Information that guides to another piece of content. In a controlled vocabulary a term pointing to another term for required or alternative usage in indexing and for prompting during a search dialogue.	e.g. synonyms, see also or related terms (RT)
Data aggregation	Inclusion or clustering models for heterogeneous data sets	
Data federation	Organized data state formed by merging and normalizing a collection of similar data objects	
Data mining	Computerized process for extracting content from structured repositories	Data mining. SEE ALSO: Text Mining
Data normalization	Standardization of identical data elements (reducing fields to the simplest meaningful or workable structure).	SEE ALSO: Normalization
Data warehouse	A central repository or information infrastructure that stores or logically connects a collection of databases and associated content with characteristics and controls that enable sharing and federated retrieval.	
Database	Repository of data organized by explicit records and fields, or tables, rows and	

	attributes	
Digital asset management	A type of content management that automates the application of rigorous governance rules for how the content is created, modified, and maintained with access controls.	DAM
Domain	A corpus of content bounded by system architecture definitions.	
Dublin core	A standard 15-element metadata element set maintained at http://dublincore.org/ as a baseline for content.	SEE ALSO: Metadata
Embedded search	Retrieval algorithms delivered as a part of a software application for searching the content within the application.	
Enterprise search	Software used to index and retrieve content that exists within or for an organization, ideally optimized for specific enterprise business requirements.	
Entity extraction	A process of content analysis by which the software identifies and classifies data by type or attribute for the purpose of creating metadata from unstructured content.	
ETL	Extract, load and transform suite of algorithms or programs	
Extractors	Software programs that harvest data content from databases, files or other applications, usually for the purpose of then manipulating the data for eventual exposure to other applications or search engines.	
Faceted navigation	In a search interface, the exposure of a controlled terminology list with facets (classes of concepts) with drill-down (broader to narrower) capabilities to facilitate moving through the facets to obtain different groups of content results.	Guided navigation
Federated search	Process of retrieving content either serially or concurrently from multiple targeted sources that are indexed separately and presenting results in a unified display.	See also Federation
Federation	Expansion of the concept of aggregation. It has play in a multi-domain environment (internal sites or a mix of internal and external). Across domains it supports at least four distinct	

	functions: ▪ Integration of the results from a number of targeted searchable domains, each with its own search engine ▪ Disambiguation of content results when similar but non-identical pieces of content might be included ▪ Normalization of search results so that content from different domains is presented similarly ▪ Consolidation of the search operation (standardizing a query to each of the target search engines) and standardizing the results so they appear to be coming from a single search operation	
Filtering	Applying other search criteria to narrow or alter the results of an existing search or stored search strategy.	SEE ALSO: Boolean searching
Full text search OR "free" text	Retrieval of strings found within the full content of a collection of files	Full text retrieval; free text search; unstructured search
Fuzzy search	Content retrieval algorithms that have rules for what content is relevant to match a query. (e.g. finding all words that are alternative grammatical forms of *elevate* or mean the same thing as *elevate*.)	
Histogram	Frequency distribution display or model – visualization presentation	
Hosted search	Retrieval software is installed and supported on computing infrastructure that is maintained by the vendor; search algorithms operate from that host on user-controlled content, which may or may not reside permanently on the host.	Often supported by host's Software as a Service or SaaS.
Index	Systematically arranged list; in computerized systems it is a representation of content to speed retrieval by the governing algorithms.	
Indexing	A human intellectual process for organizing content to optimize retrieval. A computerized process for organizing content to optimize retrieval	
Integrated information system	Connected data structures and workflow procedures with common features supporting a unified architecture and operational method.	
Interface (Search)	The architecture controlling the methods and design through which a	

	user executes a search.	
Keyword	Non-controlled terminology; language extracted from the content literally	
Keyword search	Query request for literal text as crawled and indexed by a search engine	
Knowledgebase	A domain specific data repository of facts or rules accessible in machine readable format to support software applications	
Link	URL address explicitly connecting content in one location to content in another (my be within a document, site, or remote)	Hyperlink
Loaders	Software applications designed to transfer data from one database to another often coupled with transformers	
Metadata	Explicitly defined labels for structuring content that describes any document or file regardless of the native format.	Properties, Bibliography SEE ALSO: Citation
Natural language query	Search expressed as a question by a native speaker who asks for information	NLP
Navigation	Method of traversing content with a device (e.g. mouse), or accelerator keys through a structured layer of content to reach other content (e.g. drilling down through a taxonomic structure)	
Normalization	Process or processes to create uniform format, language, and structure for data that needs to be consistently and meaningfully stored in a database and/or aggregated and federated upon retrieval.	SEE ALSO: Data federation| Federated search| Federation
OEM	Original equipment manufacturer; used to explain relationship of a supplier to another organization whose product is embedded in the delivered application.	
Ontology	An assembly of concepts in which all possible relationship that might exist between and among concepts is explicitly mapped	
Open source search engine	Retrieval software available without licensing costs and customizable by the acquiring organization or by a third-party. e.g. Lucene	
Parametric search	Interface architecture supporting the selection of multiple variable criteria in a single search pass. e.g. to find all products within a class, with specific properties, and applied to select industries.	

Personalization	Self management of the software application's interface	
Phrase search	Retrieval query specifying explicit adjacency of two or more terms in the order expressed in the query.	
Portal	Web-based page of links serving as points of entry to specific content, other web sites, and applications.	
Prompt	Interface symbol or text indicating that a user response is required to proceed with the transaction	
Repository	A database or file structure for electronic content; Entity within a searchable domain	
Results	The citations or partial content of data retrieved in a search	
Retrieval	Process of accessing content through the act of searching	
Search	Process classification for all software designed to retrieve content.	
Search appliance	Hardware bundled with search software designed to be plugged into an existing computer infrastructure to begin the process of crawling and indexing target content within a network.	
Search engine	Software with algorithms specifying how data is to be retrieved from one or more indices.	
Search intermediary	Individual who interprets what a user wants to find and performs retrieval operations on behalf of the user.	
Search platform	Suite of software products that together enhance simple index searching with additional functions related to content (e.g. transformation, analysis, reporting)	
Searching	Using retrieval software or a non-automated process for finding content	
Security	In a search environment, the search engine functions that support access controls to content through authorization validation.	
Semantic search	Use of natural language or meaningful queries to find content through retrieval software designed to understand linguistically meaningful questions and the target content.	
Site search	Option using navigation or a search box to retrieve only content from a specific Web site (URL) domain.	
Social search	Option within a search interface	

	environment to share and annotate search results using collaborative features.	
Sort	Arrange or order data in a defined sequence.	
Stemming	A form of fuzzy search logic that reduces a word to its fundamental root and looks for any word with that root. (e.g. a search for *stemming* would also retrieve *stem*, *stems*, and *stemmed*)	
Structured content	Data stored in a database or explicit metadata stored in a software application	
Structured search	Use of pre-defined forms or explicit commands to give bounds to query criteria and parameters. (e.g. restricting the search for a word to the title field)	SEE ALSO: Parametric search|Boolean search
Tag and tagging	Use for semantic labels or functional tagging that indicates the purpose of a topic or conceptual string. Different that cataloging in which metadata values are being assembled congruently to the content. Tags usually reside embedded in the content.	
Taxonomy	Hierarchically ordered list of terminology approved for tagging or categorizing a corpus of content. Also, often exposed in the search interface to form the framework for navigated search.	
Text mining	Extracting interesting and non-trivial information and knowledge from unstructured text. Interdisciplinary field that draws upon: ➢ Information retrieval ➢ Data mining ➢ Machine learning ➢ Statistics ➢ Fact extraction ➢ Computational linguistics	Information extraction, Intelligent text analysis, Text data mining, Knowledge discovery in text
Thesaurus	A list of terms that are assigned simple relationships, cross references, scope notes, usage notes and other directives. A thesaurus is often more comprehensive than a taxonomy but less complex than an ontology.	
Trackback	A URL from one piece of content to the URL of another.	
Transformers	In data and content management, tools to normalize or otherwise	

	systematically change data.	
Unstructured content	Content not organized in a formal structure; files not in a database (e.g. a Word document)	
Visualization	Graphical or image representation of data to reflect some understood relationships that reflect information or reveal knowledge about the data.	
Web search	Retrieval from a domain of content exposed to a single or multiple Web sites.	
XML	Acronym for eXtensible Markup Language. An infinity customizable markup language for defining the metatags, descriptions of kinds of content within or applied to a domain of content.	