Nair Kobashi:

- Bacharel em Jornalismo pela USP (1980);
- Bacharel em Biblioteconomia pela USP (1988);
- Mestre em Ciências da Comunicação pela USP (1988);
- Realizou estágio de pesquisa na École des Hautes Études en Sciences Socieles (1991), sob orientação de Jean-Claude Gardin;
- Professora liver-docente da USP;
- Desenvolve atividades de ensino e pesquisa na área de Ciência da Informação com ênfase em Organização, Representação e Recuperação da Informação.

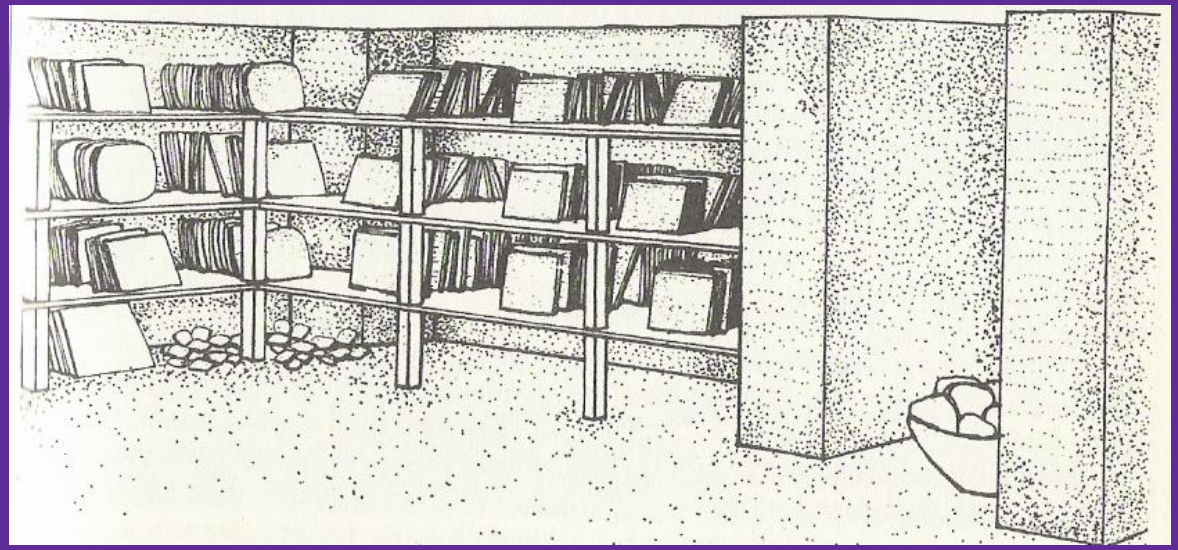
Fonte: Biblioteca Virtual da FAPESP:
http://www.bv.fapesp.br/pt/pesquisador/665928/nair-yumiko-kobashi/

Análise Documentária e Representação da Informação

Nair Yumiko Kobashi
Profª. ECA-USP

Introdução:

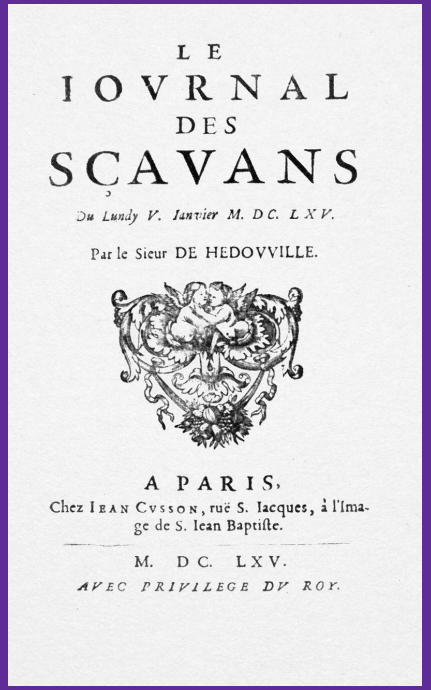
Se compararmos com as práticas de organização de documentos e formação de coleções, a preocupação teórica com a organização e a representação de informações, com finalidades documentárias: fato recente.



Biblioteca da cidade de Ebla (região da atual Síria). Em 1975 descobriram um palácio com 20000 fragmentos de tábuas de argila datadas por volta de 2.250 a.c., organizados de acordo com sua forma e conteúdo. Nesse biblioteca havia uma classificação dos materiais, as tábuas tinham uma espécie de escrita nas laterais para facilitar sua localização.

Nas cidades-estado da Mesopotâmia, “as obras cunhadas em tábuas de argila eram protegidas por espécie de envelopes, sobre os quais se transcreviam informações que cumpriam função semelahnate aos resumos” (p.5)

Documentação enquanto sistema público de informação:



LE
IOVRNAL
DES
SÇAVANS

Du Lundy V. Ianvier M. DC. LXV.

Par le Sieur DE HEDOVVILLE.

A PARIS,
Chez IEAN CVSSON, ruë S. Iacques, à l'Image de S. Iean Baptiste.

M. DC. LXV.

AVEC PRIVILEGE DV ROY.

"Le Journal des Sçavans"
Publicação: Paris, 1666, pela Academia Francesa de Ciências.
Apresentava resumos dos trabalhos cientificos, filosóficos e artísticos produzidos no período.

jornal de Resenhas

NELSON SIERA DO NASCIMENTO (ORG.)

VOL. 1

Publicações semelhantes, em suporte de papel, cresceram ao longo dos séculos, até chegarem nos formatos atuais: fitas magnéticas, disquetes, CD-ROMs, etc..

Década de 1950:
marco da sistematização das reflexões teóricas sobre os processos de documentação.

Motivação:

Experimentos automáticos de indexação, elaboração de resumos e recuperação de informação.

Luhn (1958-1959):

Critérios estatísticos de ocorrências de frequência.

Durante quase duas décadas sua contribuição “foi considerado o paradigma norteador de inúmeros projetos que visavam modificar radicalmente os métodos tradicionais de armazenamento e, tratamento e recuperação de informação” (p.6).

Limites práticos dos modelos baseados no paradigma de Luhn:

São eficientes para identificar formas significantes, mas pouco útil para determinar significados.

FABRICIO SHIGUERU CATAE

CLASSIFICAÇÃO AUTOMÁTICA DE TEXTO
POR MEIO DE SIMILARIDADE DE PALAVRAS:
UM ALGORITMO MAIS EFICIENTE

Dissertação apresentada à Escola Politécnica
da Universidade de São Paulo para obtenção
do Título de Mestre em Engenharia Elétrica.

São Paulo
2012

FICHA CATALOGRÁFICA

Catae, Fabricio Shigueru
Classificação automática de texto por meio de similaridade de palavras: um algoritmo mais eficiente / F.S. Catae. -- São Paulo, 2012.
93 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1. Algoritmos 2. Classificação automática de texto 3. Reconhecimento de padrões I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II. t.

A Interface Linguística - Análise Documentária

Novas hipóteses de Trabalho:

Valorização dos aspéctos comunicacionais dos Sistemas Documentários e aos instrumentos que possam promover níveis crescentes de interação entre sistema e usuários.
Exemplo: Tesauros, passam a ser vistos como instrumento para auxílio aos usuários.

Tesauros:

“Vocabulário controlado e dinâmico de descritores relacionados semântica e genericamente, que cobrem um específico de conhecimento”;

“Lista estruturada de termos associados de informação e indexadores, para descrever um documento de modo conciso e com a especificidade necessária, na etapa de entrada em sistemas de informação e também na etapa de recuperação posterior”.

(fonte: CUNHA, Murilo e CAVALCANTI, Cordélia. Dicionário de Biblioteconomia e Arquivologia, p.368)

Aproximação:

Análise Documentária e Linguística Aplicada.

Justificativa inicial: semelhanças entre os processos documentários e a tradução automática

Destaque: arqueólogo Jean-Claude Gardin

Criou a denominação do campo que hoje se conhece por Análise Documental.

Criou as bases fundantes de uma teoria específica sobre os processos documentários.

Interesse de Gardin:

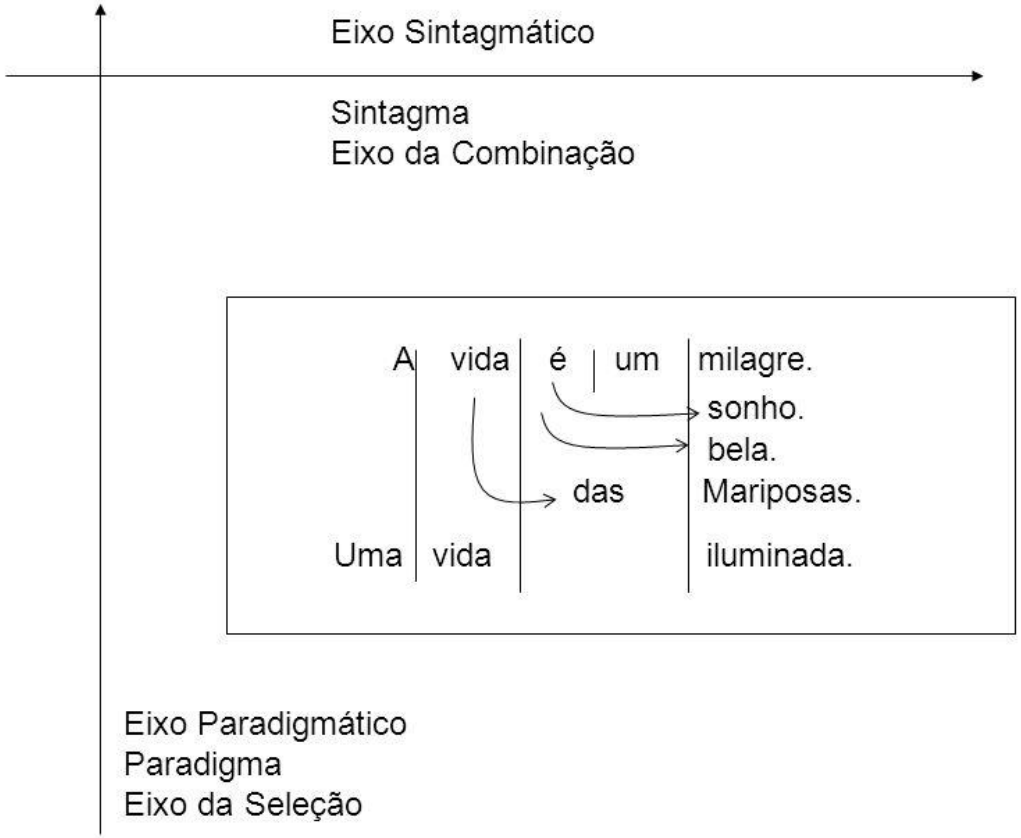
Compreender os aspectos documentários como processos relacionados à passagem do texto original para a sua representação

através de linguagens documentárias.

Propôs um modelo formal de linguagem documentária.

Linguagem Documentária é constituída por dois “elementos” principais:

1) conjunto de termos (o léxico), por relações entre as unidades lexicais, determinadas a priori (o eixo paradigmático);
2) por uma sintaxe que articula os encadeamentos entre termos da linguagem (eixo sintagmático).


Eixo Sintagmático
Sintagma
Eixo da Combinação
A vida é um milagre.
sonho.
bela.
das Mariposas.
Uma vida iluminada.
Eixo Paradigmático
Paradigma
Eixo da Seleção

Montgomery (1972)
Reforça a importância da aproximação entre Linguística e Ciência da Informação,

para o benefício de ambas as áreas.

“A mais recente produção teórica (...) aponta para interface Terminologia-Análise Documentária como um campo com possibilidades de imprimir rigor às práticas empíricas de construção de vocabulários para fins documentários” (p.7)

A organização e representação das informações documentárias

A Análise Documentária, em sua dimensão operatória, manipula e transforma textos em dois tipos básicos de representações: resumo e índice.

Visão simplista:

- **Resumos:** atos técnicos de reduzir textos;
- **Indexação:** estabelecer equivalência entre palavras do texto e uma linguagem documentária.

Conceito de Comunicação Documentária:

- Ato documentário é um ato de comunicação, cuja finalidade é promover a circulação da informação;

- Deve seguir metodologias (ler o texto, selecionar conteúdo pertinente e criar uma representação);

- Deve seguir parâmetros específicos: objetivos institucionais, tipos de usuários, demandas, estrutura organizacional, etc.

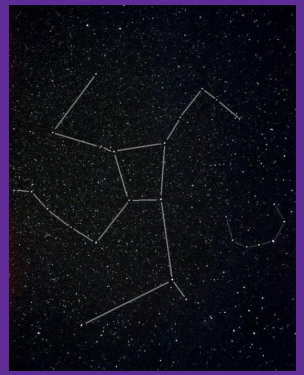
Linguagem Natural

- Palavras podem ser imprecisas, subjetivas;
- Significados múltiplos;
- “A natureza genérica das palavras é um fenômeno geral da linguagem, um elemento de abstração da linguagem” (STEPHEN ULLMANN)

Linguagem Documentária

- Termos voltados para a univocidade (sentido inequívoco);
- Conteúdos prefixados;
- Representam “classes” e não “informações particulares”.

Libra?



£5
Bank of England
5
FIVE
Pounds
£10
10
TEN
£20
20
TWENTY
£50
FIFTY

ALFABETO DE LIBRAS
A B C Ç D E F G H I
J K L M N O P Q R
S T U V W X Y Z

SP
RJ
São Paulo
Rio de Janeiro
Santos
Ilha Bela
Libra
Iara
Tupi e Iracema
Tupi
Guará
Fonte: Petrobras

No âmbito pragmático da recuperação da informação, a indexação e o resumo atuam de forma complementar

O proceso global de Análise Documentárua defronta-se com pelo menos dois objetos:

O Texto, como a sua unidade de análise;

A Linguagem Documentária, como instrumento comutador.

Como abordar tais objetos?

Quais são os procedimentos necessários para se observar o documento?

Que tipo de técnicas podem otimizar o processo de fabricação das representações documentárias?

A Representação Documentária

O que é REPRESENTAÇÃO?

-Reprodução daquilo que se pensa;

-Substituição;

-Descrição;

-Reproduzir algo;

-Descrever;

-Tornar algo presente;

-Interpretrar.

Significados próprios para:

- A filosofia;
- A linguística;
- A ciência da comunicação;
- A semiótica;
- Etc.

A Representação Documentária

REPRESENTAÇÃO na Documentação:

Adequação dos temos para a linguagem da indexação.

Um conceito associado a:

Catalogação - Descrição de aspectos que identifiquem materialmente os documentos;

Indexação e Elaboração de Resumos - Processo e produto da condensação de conteúdos de texto.

A Representação Documentária

INFORMAÇÃO DOCUMENTÁRIA

Uma representação que deve procurar responder às exigências impostas pela recuperação da informação, mantendo uma relação de similaridade com o texto original.

Obtida através da neutralização do poder expressivo do texto;
Moldada segundo regras previamente determinadas;
Apesar de ser uma ”representação” (diferente do original), deve ser equivalente a ele, do ponto de vista informacional.

A Significação dos Processos Documentários

Conceito de SIGNIFICAÇÃO:

Diálogo entre Documentação e Ciências da Linguagem (tal como proposto pela Semiologia).

Semântica do Código: permite discutir as relações de significação entre as unidades construtivas das linguagens documentárias;

Semântica Textual: conceitos para se operar com as apropriações parafrásticas que caracterizam o resumo.

A Significação dos Processos Documentários

Divergência dentro da Semiologia:

x

Semiótica de Peirce: conceito fundamental é a **simiose** (processo pelo qual qualquer coisa age como signo – isto é, significa);

Semiótica de Saussarre: o objeto da Linguística, enquanto parte da Semiologia, é a **língua**.

A Significação dos Processos Documentários

A Linguagem Documentária (Linguagem de Indexação, Sistema de Classificação, Linguagem de Informação, Lista de Cabeçalhos de Assuntos…) é um código normalmente composto por um conjunto limitado de termos, que prescreve formas de entrada e de busca a serem utilizados pelo indexador ou pelo usuário.

Suas unidades são selecionadas de diversas fontes (Linguagem de Especialidade, Linguagem de Uso Corrente, Terminologia da Área…). Assim, apresentam questões semânticas complexas para a sua estruturação (significação em que as relações entre signos e referentes são de natureza bastante diversa).

A Significação dos Processos Documentários

Terminologias são instrumentos que atuam como fonte de referência na construção da Linguagem Documentária.

São próprias de campos de conhecimento especializado.

Assim, a Linguagem Documentária não substitui as Terminologias (Linguagem Especializada) em sua função comunicativa geral. Ela atua na tarefa de recuperar informação.

Análise do discurso e a Linguística do texto

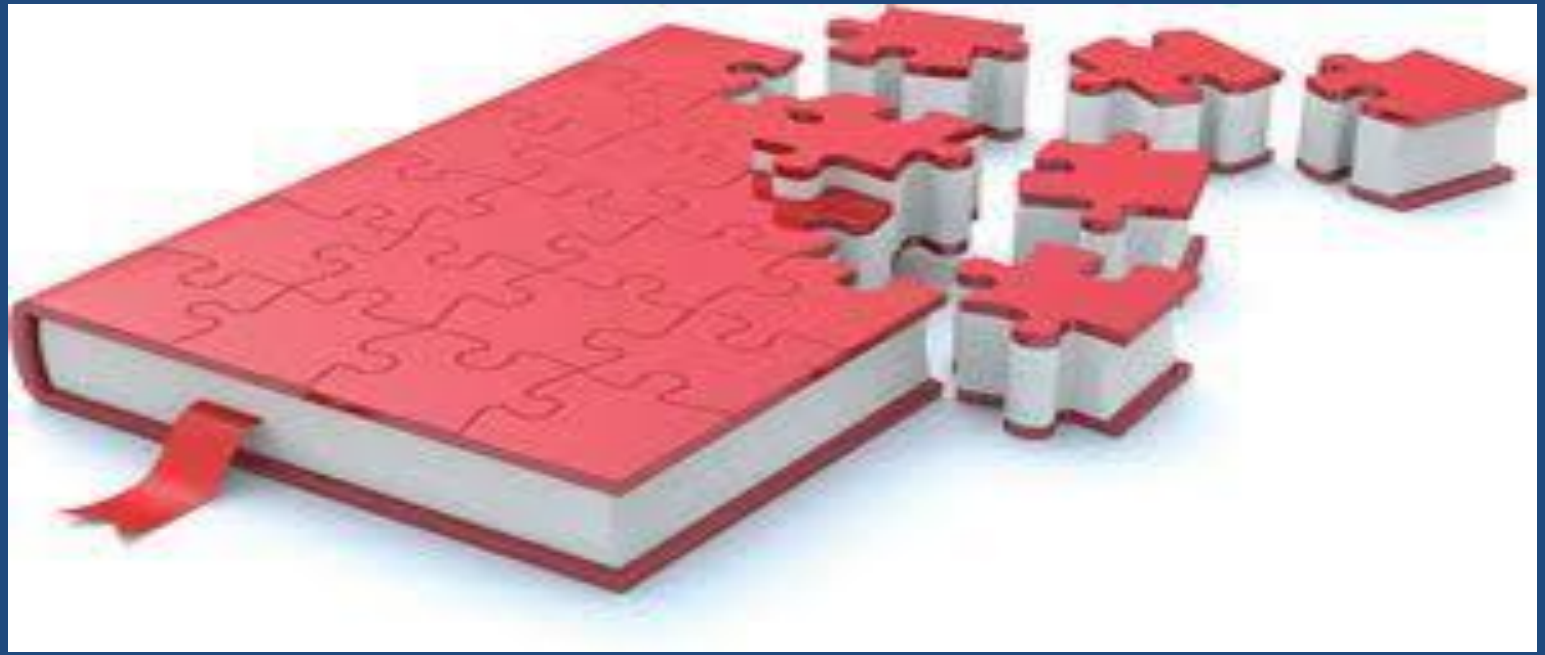
Michel Pêcheux

- "integrar, por meio da enunciação, uma abordagem interna do texto, indispensável para que se reconheçam os mecanismos e regras de engendramento do discurso, com análise externa do contexto sócio histórico em que o texto se insere e de que, em última instância, cobra sentido." (BARROS, D. L.P. 1988. p. 5)

Entrave na análise do discurso

Linguística do Texto

Superestrutura



Análise logicista

- "... a expressão concreta de um raciocínio científico é o texto científico, no qual o autor expõe as operações do espírito que o conduziram, da observação de certos fatos empíricos, ao enunciado de proposições denominadas, deforma diversa: teses, hipóteses, interpretações, comentários, conclusões, explicações, etc." (GARDIN, J. C. 1987a. p. 4)

Representação da análise logicista

Análise logicista X Análise documentária

a Análise de Texto na perspectiva da Inteligência Artificial

o que é IA?

área interdisciplinar voltada à construção de sistemas com características associadas à inteligência do comportamento humano

aquisição de conhecimentos, raciocínios e processos cognitivos

compreensão da linguagem natural

aplicações

reconhecimento de imagens, formas e voz

automação de tarefas

IFTTT

tradução automática

história da IA

1936	1950	anos 80
primeiras iniciativas	campo autônomo	consolidação da área

interdisciplinaridade

Informática
Psicologia cognitiva
Linguística
Lógica

hipótese fundamental

os processos de pensamento são **mecanizáveis**, simuláveis em computador

2

questões fundamentais

para o tratamento automático das línguas

(Gerard Sabah, 1990)

o que significa
compreender
a linguagem?

como um programa pode reconhecer algo que não tem sentido ou possui sentido ambíguo?

roubaram
o cachorro
do meu tio...

um programa deve ser capaz de identificar os objetivos do interlocutor e inferir seu plano para interpretar corretamente o que é dito

compreender a representação interna da frase

a compreensão automática de um texto prevê a construção da representação interna

operações automáticas de tradução

elaboração de resumos

indexação

recuperação de informação

exemplos

projeto TOPIC

Text Oriented Procedures for Information Management and Condensation of Expository Texts

Universidade de Konstanz (Alemanha)

condensa textos de 2000 a 4000 palavras

módulo de indexação automática (vocabulário controlado)

INIST

Institut de L'information Scientifique et Technique

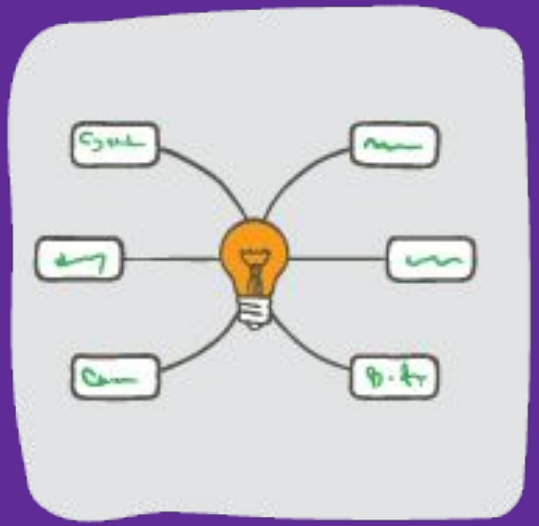
construção de banco de dados (Pascal | Francis)

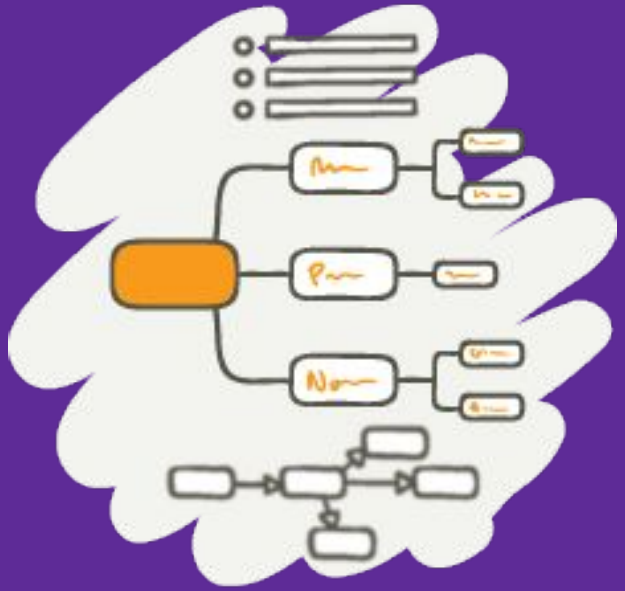
tratamento linguístico e estatístico dos termos

resultados de buscas em forma de mapas conceituais

INIST

mapas conceituais





Juan Barceló

necessidade de estruturação do conhecimento - CNRS (J.Claude Gardin)

sistema de produção de dados novos a partir da articulação de conhecimentos

3 tipos de metaconhecimentos sobre:

problema a ser resolvido
estrutura do sistema
estratégia de resolução

sistemas especialistas

perguntas feitas ao usuário para encontrar a resposta final

P: Você sabe o nome do autor que está procurando?”

R: Não

P: Você gosta de romances policiais?

R: Não sei...

exemplos de ferramentas

para construção de sistemas especialistas

Expert Sinta
Universidade Federal do Ceará

Prolog
linguagem de programação associada à IA e linguística computacional

CLIPS
desenvolvido pela NASA

otimismo

cauteloso

Considerações Finais

Análise Documental

Balizou as reflexões da autora sobre questões relacionalidades à Organização e Representação Informações Documentárias.

A autora constatou a necessidade de discutir os problemas de Representação e de Significação. Razão pela qual revisitou diferentes conceitos construídos pelas Ciências da Linguagem.

Qual a singularidade da Análise Documentária frente os demais tipos de análises de textos?

Apesar de ter, como em outros tipos de análises, o objeto-texto como foco de preocupação, na Análise Documentária “os textos são submetidos a um conjunto de operações que visam transformá-los em produtos documentários” (p.22).

Apesar de ter, como em outros tipos de análises, o objeto-texto como foco de preocupação, na Análise Documentária “os textos são submetidos a um conjunto de operações que visam transformá-los em produtos documentários” (p.22).

Diferenças da AD com as demais áreas do conhecimento:

Análise do Discurso: propõe gerar interpretações singulares;
Análise Logistica: tem como objetivos mapear e classificar raciocínios;
Línguística de Texto: propõe a identificar fenômenos linguísticos;
Análise de Conteúdo: propõe a identificar as condições

de produção de texto.

Objetivos da Análise Documentária:

Se propõe a elaborar diversas modalidades de informações, quer sob a forma de novos textos (resumos) ou de representações padronizadas, mediadas por códigos de convenções, ditos vocabulários controlados" (p.22), visando a circulação de informação nas várias esferas da atividade humana.

Para Reflexão:

“O Rigor da Ciência”,
de Jorge Luis Borges.

«Sobre o Rigor na Ciência

…Naquele império, a Arte da Cartografia alcançou tal Perfeição que o mapa de uma única Província ocupava uma cidade inteira, e o mapa do Império uma Província inteira. Com o tempo, esses Mapas Desmedidos não bastaram e os Colégios de Cartógrafos levantaram um Mapa do Império que tinha o Tamanho do Império e coincidia com ele ponto por ponto. Menos Dedicadas ao Estudo da Cartografia, as gerações seguintes decidiram que esse dilatado Mapa era Inútil e não sem Impiedade entregaram-no às Inclemências do sol e dos Invernos. Nos Desertos do Oeste perduram despedaçadas Ruínas do Mapa habitadas por Animais e por Mendigos; em todo o País não há outra relíquia das Disciplinas Geográficas.

(Suárez Miranda: Viajes de Varones Prudentes, libro cuarto, capítulo XIV, Lérida, 1658.) »