A busca pela eficiência na representação da informação e do conhecimento – desdobramentos posteriores no pensamento de Gardin

Johanna W. Smit
Universidade de São Paulo
johannawsmit@gmail.com

1 Introdução

A relação entre o texto original e sua representação acompanhou Gardin ao longo de toda sua trajetória, e ele a discutiu não somente pensando em sistemas de informação, mas nas diferentes modalidades de análise de textos.

Este não é o momento para detalhar uma bibliografia importante na qual Gardin discutiu os diferentes métodos de análise de textos, mas uma rápida referência à questão não pode ser ignorada, tendo em vista que a análise documentária (AD), em sua visão, representa uma modalidade de análise de textos – segundo ele, o ramo mais "industrial" das análises (Gardin, 1974, p.45). Enfatizando a dimensão semântica das análises, ele elaborou uma visão comparativa dos procedimentos adotados, por exemplo, pela análise de conteúdo exercida por sociólogos e outros profissionais, a análise estrutural de contos e narrativas, a análise semiológica ou literária e a análise documentária, ou seja, "a expressão do conteúdo de textos científicos à maneira dos documentalistas" (Gardin, 1970, p.630). Estas comparações o levaram a postular a importância da explicitação do método de análise de textos (em sua formulação: como passar de T1, original, a T2, sua representação), a explicitação das correspondências semânticas mobilizadas pela análise e a possibilidade de verificação da relação entre o texto original (T1) e sua representação (T2). Dito em outros termos, ele propôs discussões acerca dos **métodos de análise** (e sua explicitação), das **ferramentas** (ou seja, a metalinguagem usada para extrair o conteúdo semântico dos textos, de

acordo com os objetivos propostos) e estratégias de **avaliação** e **validação** das análises.

Inserir a análise documentária num conjunto maior de análises de textos representou, certamente, uma ampliação do conhecimento na documentação e biblioteconomia, ao chamar a atenção para questões – linguísticas - até então consideradas resolvidas pelo bom-senso. Entre a atribuição de um código de classificação a um livro e as discussões sobre a unidade de análise (palavra? frase? parágrafo? capítulo? o texto inteiro?), a preponderância da semântica sobre a sintaxe e o potencial de automatização do processo de classificação, elaboração de resumos ou indexação separam o bom-senso de uma visão que se quer mais científica e rigorosa da prática biblioteconômica ou documentária.

A importância da metalinguagem (ML) usada para analisar os textos - a linguagem documentária (LD) na Ciência da Informação - recebeu muita atenção e suponho que esta questão seja a que mais foi incorporada pela Ciência da Informação. Uma linguagem para representar o conteúdo original dos textos deve ser uma linguagem artificial, construída de acordo com objetivos e, portanto, distinta da linguagem natural. A origem da temática pode ser encontrada quando o arqueólogo Gardin discutiu como encontrar correspondências (ou seja, a classificação) entre objetos arqueológicos, ou seja, quais critérios adotar para reunir objetos em conjuntos, distinguindo-os de outros conjuntos. Nas décadas de 50 e 60 ele elabora “códigos analíticos” de cilindros orientais, artefatos, desenhos decorativos, etc. enfatizando sempre a necessidade de comparações para estabelecer relações (Gardin, 1979, p.87).

A preocupação com a importância do estabelecimento de relações entre dados e informações fica clara também, quando ele faz parte da equipe que discute a concepção do UNISIST na Unesco, em 1970, ao afirmar que “não adianta disponibilizar os dados, as pessoas têm que conseguir entender os dados, estabelecendo as relações entre os dados”.

É verdade que a tecnologia da época nos parece hoje bem rudimentar, pois a seleção de objetos que apresentavam características comuns foi concebida em fichas com perfuração lateral (peek-a-boo). Uma análise dos conceitos presentes no Alcorão também deu margem à elaboração de um vocabulário, estruturado, de conceitos, e testes feitos com estas fichas (Allard et al, 1963).

O investimento em linguística foi muito grande (Gardin, 1973): consideramos que hoje o Grupo TEMMA deu prosseguimento a este investimento, mas com ênfase na terminologia, visando questões relacionadas ao controle de vocabulário e o poder comunicacional dos produtos da análise documentária.

A necessidade de teorização das linguagens documentárias, ou seja, as metalinguagens específicas dos ambientes de tratamento da informação, levou Gardin a discutir suas estruturas e sua composição, distinguindo nelas um **léxico, relações paradigmáticas** e **relações sintagmáticas** (Gardin, 1966, 1970, 1973). Ou seja, aquilo que até aquele momento muitos faziam baseados no bom-senso, passou a contar com uma teoria, fortemente ancorada na linguística.

Em relação ao **léxico**, ou seja, os termos que compõem a linguagem documentária, deve-se frisar a necessidade de seu controle, eliminando a sinonímia, resolvendo a homonímia, etc.

Em relação às **relações paradigmáticas -** principalmente as relações hierárquicas reconhecidas e estabelecidas pelos diferentes domínios do conhecimento - seu maior interesse reside na afirmação de que as mesmas não representam automaticamente a organização de determinado universo do conhecimento, mas os objetivos do serviço de informação, o que reforça o caráter artificial das linguagens documentárias.

No entanto, desde a década de 60, Gardin se preocupava com as relações que podem ser estabelecidas entre termos do léxico, mas que não configuram uma relação paradigmática, ou seja, as **relações sintagmáticas**, chegando ao ponto de codificá-las quando do desenvolvimento do SYNTOL (Syntagmatic Organization Language), como será detalhado a seguir.

A otimização da relação entre informação original (T1) e informação representada (T2) pode ser analisada, nas propostas de Gardin, por três desdobramentos:

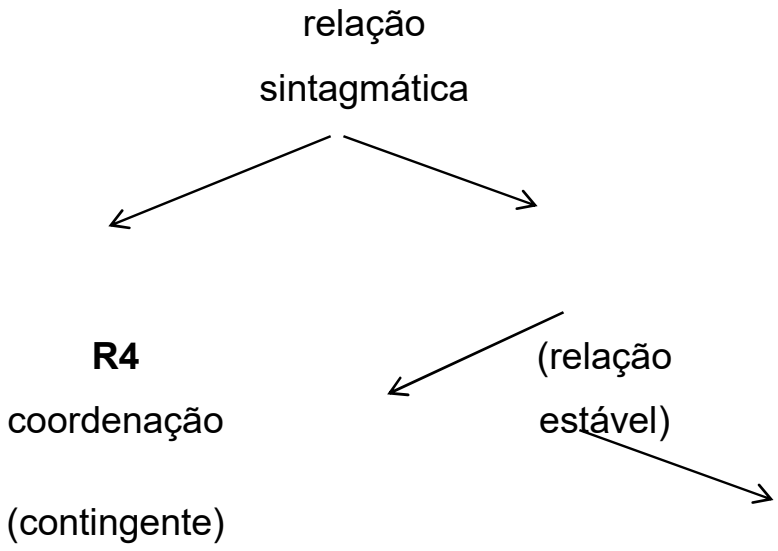
1. Propor a construção de ontologias (cf. seção 1);
2. Adotar uma visão extremamente positivista da ciência (cf. seção 2) e
3. Face ao custo envolvido nas propostas que serão apresentadas e discutidas nas seções 1 e 2, ele imaginou outra abordagem, mais radical, e que propunha a reformulação do texto científico, visando torná-lo mais consultável (cf. seção 3).

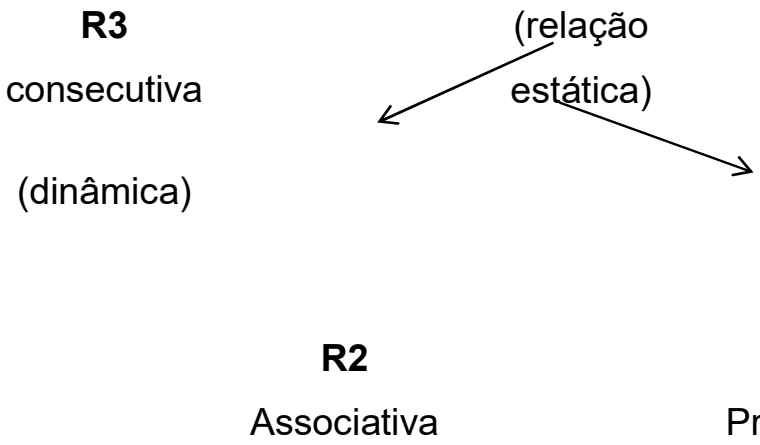
1 A proposta de ontologias antes da hora

O SYNTOL foi desenvolvido na França, por um grupo coordenado por Gardin, por encomenda da Euratom (European Atomic Energy Community) entre 1960 e 1962. O SYNTOL não é uma LD mas um sistema geral de documentação automática, objetivando a indexação automática de textos, composto de um léxico organizado (ou seja, termos controlados e organizados paradigmaticamente) e um sistema bastante elaborado de relações sintagmáticas, para estabelecer a interconexão entre os termos no caso de assuntos mais complexos. "O SYNTOL é um conjunto de regras lógico-linguísticas, a respeito das diferentes maneiras de representar as informações que se encontram nos documentos científicos, com vistas a sua exploração pelo computador" (Cros, Gardin, Levy, 1964, p.39-40).

Extremamente detalhado e muito complexo, a concepção do SYNTOL é muito interessante do ponto de vista teórico, mas pouco usado na prática. Testes foram realizados em 1962, para textos das áreas de fisiologia, psicologia, sociologia e etnologia e permitiram concluir que o Syntol, nas palavras de Gardin, era "muito canhão para pouca mosca".

Se a implantação prática ficou comprometidao (por conta tanto da complexidade do sistema como das limitações tecnológicas da época), importa reter da experiência o caráter precursor da proposta de padronização de relações entre os termos no momento da indexação. As relações sintagmáticas foram organizadas e codificadas em conjuntos lógicos (donde a expressão OLS - Organização Lógico-Semântica) e apresentadas na forma de diferentes árvores de relações. Por exemplo:





A relação entre “álcool” e “alegria”, por exemplo, produzia o sintagma <álcool R3 alegria>, onde R3 denotava uma relação consecutiva. Deve-se salientar que entre as décadas de 70 e 90 o recurso aos *role operators,* ou marcadores, foi muito utilizado. Trabalhando com textos na área médica, por exemplo, passava a ser importante poder distinguir uma droga na condição de medicamento para combater certa doença ou na condição de provocador de uma doença. Hoje parece simples associar o que foi proposto àquela época às atuais ontologias, mas seu caráter precursor, com uma tecnologia ainda incipiente, não pode ser ignorado.

A formalização das relações levou à definição do sintagma mínimo de qualquer representação da informação: Ri (x,y), onde *R* representa uma relação (paradigmática ou sintagmática), *x* e *y* são termos do léxico. A partir da multiplicação de LDs e da elaboração de tesauros multilíngues, a noção do sintagma mínimo mostrou seu potencial nos estudos de compatibilidade e interoperabilidade entre LDs.

Deve-se reter destas propostas, datada da década de 60, a evidente conexão com os princípios da inteligência artificial, também em amplo desenvolvimento naquela época, através da explicitação de relações entre os termos e análise de condições nas quais as relações podem ser estabelecidas. A mesma temática é hoje atualizada no desenvolvimento de ontologias, ao identificar “objetos”, propriedades, regras de relacionamentos e condições de aplicabilidade: este conjunto compõe a base de conhecimento e o motor de inferência. (Gardin, 1989, p.16-17).

2 A avaliação da representação da informação

A capacidade representacional dos produtos da AD recebeu muito atenção por parte de Gardin e certamente representa um veio de pesquisa ainda pouco explorado pelo Grupo TEMMA. Esta avaliação do produto da AD pode ser sistematizada a partir de dois pontos de vista, complementares mas totalmente distintos:

- a avaliação da "satisfação" do usuário a partir de uma busca em sistema de informação. Cálculos de precisão, revocação, ruído, etc. foram muito desenvolvidos pela bibliografia e são, evidentemente, muito importantes;
- mas Gardin considerava outro cálculo igualmente importante e que dizia respeito ao grau de representatividade do produto da análise em relação aos textos originais (Gardin et al, 1981).

As duas avaliações não devem ser consideradas auto-excludentes, mas complementares, embora a segunda avaliação, que enfatiza a representatividade da representação em relação ao texto original não seja comum em nossa bibliografia.

Sobrepõe-se, nesta questão, a meu ver, a figura do cientista àquela do documentalista. É interessante observar que a representatividade do texto 2, produzido a partir do texto 1 é na prática, sempre pressuposta, mas não discutida. Gardin não transformava o tema em uma questão ética ou moral mas numa questão prática: ele considerava que o usuário do sistema de informação para poder, efetivamente, avaliar o resultado de suas buscas no sistema de informação, devia ter acesso a toda informação relacionada aos procedimentos de AD adotados, bem como todos os detalhes da LD empregada. Somente o conhecimento de todos os detalhes que integram o processamento da informação faz com que o usuário possa sair de uma posição passiva – receber o resultado de uma busca – para uma posição ativa de real avaliação do resultado da busca efetuada.

Um usuário em condição de avaliar, autonomamente, o resultado de uma busca e, portanto, validá-lo: este é um desafio lançado por Gardin que, ao que me parece, ainda está longe de ser respondido. Mas o desafio é bonito, instigador, e merece integrar a agenda de pesquisa. A satisfação do usuário, neste caso, se completa por uma recuperação de informações que não somente corresponde aos termos utilizados em sua pesquisa mas que remetem a textos que igualmente correspondem ao que foi buscado.

3 O projeto logicista e a re-escrita do texto científico

Naquilo que me parece ser a última fase das empreitadas de Gardin, ele se interroga cada vez mais sobre o sentido social das publicações científicas e dos sistemas de informação. Na década de 80 e parte da década de 90 ele investiu muito na temática da análise de textos, em diferentes áreas, tentando identificar a consistência das mesmas: ao final ele considerou demonstrado que os textos nas áreas das ciências sociais e humanas apresentavam muitos vícios, recheavam publicações e bibliotecas mas que não era possível perceber o avanço do conhecimento. Obviamente esta conclusão é discutível, mas faz sentido quando colocada na perspectiva gardiniana da total explicitação dos procedimentos de representação da informação. Inconformado com este estado de coisas, ele elabora novas propostas, alicerçadas em todo o trabalho anterior, objetivando uma solução mais lógica, ou útil, para a produção do conhecimento em ciências sociais e humanas.

Reaproximando-se novamente da questão da documentação, ele retoma, sem citar, a dualidade enunciada por Briet, ao identificar uma ordem primária dos textos, produto da atividade científica e técnica que é apresentada de acordo com hábitos estabelecidos pelas diferentes áreas do conhecimento e uma ordem secundária, na qual ocorre a produção da informação documentária e criação de acessos à ordem primária. Gardin considera que a documentação se tornou refém da ordem primária, tendo em vista que a produção de textos científicos aumentou exponencialmente.

Em seguida, ele postula estar ocorrendo uma “crise das publicações”, em função de um desequilíbrio flagrante entre a produção de textos científicos e a capacidade humana de consumi-los. A criação de ferramentas para acesso aos documentos originais - pela melhoria de procedimentos de seleção de textos através da indexação, elaboração de resumos, sistemas de alerta, mineração de dados e etc. - evidentemente amenizam o desequilíbrio entre produção e consumo, mas não resolvem o problema, até porque a capacidade de consumo de informação, pela mente humana, mudou muito pouco nos últimos séculos, face ao crescimento exponencial da publicização de textos. “A busca incessante de otimização entre silêncios e ruídos” permanece, mesmo quando amenizada (Gardin, 2001, p.2).

A próxima pergunta que Gardin formula é mais provocante: em supondo um sistema de documentação totalmente adequado, que não elimine documentos pertinentes em resposta a uma consulta, aliado a uma seleção bem calibrada de textos pertinentes, pode-se postular que todos deveriam ser lidos, pois pertinentes, mas que

certamente somente uma parte será lida integralmente e que muitos serão simplesmente consultados. Ele cita experimentos feitos de seleção de textos em determinada área, por especialistas para atestar sua pertinência, e cálculos do tempo necessário para sua leitura, concluindo pela total impossibilidade de leitura de todos os textos considerados pertinentes. A partir desta constatação, Gardin propõe a remodelagem dos textos científicos de forma a que possa sem consultados, tendo em vista que não serão lidos. A “crise das publicações científicas” levanta ainda outra questão: quantos leitores, em média, terá um artigo científico? “É razoável manter um sistema de publicação de artigos cujos leitores são tão raros”? (Gardin, 2001, p.3).

A partir destas interrogações Gardin estima que o problema não está em tornar acessível os melhores textos (isto a documentação sabe fazer e faz muito bem), mas em conseguir consumi-los. Pode-se, portanto, reformular a afirmação de Bush, quando este propunha “a tarefa massiva de tornar mais acessível um desconcertante acervo de conhecimentos” (Bush, 1945, p.1) e afirmar que não é mais possível continuar escrevendo textos para serem lidos, quando a maior parte será somente consultada. Enquanto os textos científicos não forem remodelados, a documentação continuará refém do sistema de publicações científicas.

Trata-se, portanto, de uma proposta de intervenção na produção de publicações, decompondo-as em diferentes camadas:

- fatos e dados – vestígios materiais, acontecimentos históricos, práticas culturais;
- teorias, ou seja, pontos de vista formulados sobre os fatos ou dados;
- os procedimentos adotados para passar dos fatos às interpretações e teorias;
- os exemplos citados para corroborar a elaboração de interpretações e teorias.
- Percebe-se nesta proposta a volta dos conceitos da base de conhecimento e do motor de inferência, caros aos sistemas especialistas. As camadas, hipertextuais, devem, em sua visão (Gardin, 2001):
- preservar o conteúdo informacional dos textos;
- explicitar a base de conhecimento e as regras de passagem de T1 para T2 (regras de raciocínio);

- excluir as digressões metodológicas, recursos retóricos, referências filosóficas, charmes literários, etc.

A proposta foi concretizada ao remodelar um livro de 500 páginas - sobre as pérolas de cornalina, recolhidas em sítios arqueológicos da Índia e do Oriente Próximo e que levaram a inferências sobre as sociedades que as fabricavam ou importavam - em um CD-ROM multimídia contendo o mesmo conteúdo informacional, com navegação hipertextual, separando e relacionando fatos, ilustrações, argumentos de apoio, conclusões, esquematizações dos procedimentos adotados para passar dos fatos às conclusões, etc.

A categorização de tipos de informação é seguramente uma competência desenvolvida pela Ciência da Informação, razão pela qual Gardin propõe um trabalho cooperativo entre profissionais de informação, editores e autores, reconhecendo que a clareza acerca de léxicos e relações entre termos é apanágio dos profissionais de informação.

4 Em resumo

Além das contribuições de Gardin ao introduzir a linguística nos estudos de classificação, AD e indexação, restam-nos os desafios por ele lançados:

- reconhecer a documentação como uma etapa linguística do fazer científico, ao estabelecer terminologia e discutir as relações entre termos, de acordo com os campos do conhecimento;
- desenvolver sistemas de avaliação dos produtos de nosso trabalho que aliem os estudos de satisfação dos usuários a avaliações da representatividade de nossos produtos em relação aos textos representados;
- investir nas ontologias como forma de contribuição à indústria de publicações de textos científicos, para torna-los consultáveis. Neste esforço os dados inciais e as regras de inferência devem ser totalmente explicitados: a disponibilização de dados abertos já é muito discutida hoje, mas a explicitação das regras de inferência , que supõem a repetibilidade nas ciências, ainda é um tabu nas ciências sociais e humanas.

Referências

ALLARD, M., ELZIERE, M., GARDIN, J.-C., HOURS, F.(1963). Analyse conceptuelle du Coran sur cartes perforées. Paris: Mouton.

BUSH, V. (1945). As we may think. Atlantic Monthly, julho.

CROS, R.C., GARDIN, J.-C., LEVY, F. (1964). Le Syntol, um système général de documentation automatique. Paris: Gauthier Villars, 1964.

GARDIN, J.-C. (1989). Artificial intelligence and the future of semiotics: an archaeological perspective. Semiotica, vol.77, n.1/3, p.5-26.

GARDIN, J.-C. (1973). Document analysis and linguistic theory. Journal of Documentation, vol. 29, n.2, p.137-168.

GARDIN, J.-C. (1966). Eléments d'um modéle pour la description des lexiques documentaires. Bulletin des Bibliothèques de France, n.5, p.171-182.

GARDIN, J.C. (1974). Les analyses de discours. Neuchâtel: Delachaux et Niestlé.

GARDIN, J.-C. (1970). Procédures d'analyse sémantique dans les sciences humaines. In: POUILLON, J., MARANDA, P. Échanges et communications: mélanges offerts à Claude Lévi-Strauss à l'occasion de son 60ème anniversaire. Paris, Mouton, p.628-657.

GARDIN, J.-C. (1979). Une archéologie théorique. Paris: Hachette, 1979.
GARDIN, J.C. (2001). Vers um remodelage des publications savantes: ses rapports avec les sciences de l'information. In: CHAUDIRON, S., FLUHR, C. Filtrage et résumé automatique de l'information sur les réseaux: actes du Chapitre Français de l'ISKO. Nanterre: Université de Paris X, 2001, p. 3-8.

GARDIN, J.-C. et al. (1981). La logique du plausible: essais d'épistémologie pratique em scieces humaines. Paris: Editions de la Maison des Sciences de l'Homme.