

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE BOARD OF PATENT APPEALS
AND INTERFERENCES

Ex parte THOMAS M. BREUEL

Appeal 2007-2627
Application 10/248,681
Technology Center 2100

Decided: March 14, 2008

Before LANCE LEONARD BARRY, JEAN R. HOMERE, and JAY P. LUCAS, *Administrative Patent Judges*.

BARRY, *Administrative Patent Judge*.

DECISION ON APPEAL

I. STATEMENT OF THE CASE

A Patent Examiner rejected claims 1-30. The Appellant appeals therefrom under 35 U.S.C. § 134(a). We have jurisdiction under 35 U.S.C. § 6(b).

A. INVENTION

The invention at issue on appeal automatically extracts information from web pages. The Appellant observes that structured information formatted according to the Hypertext Markup Language ("HTML") is common on the Internet. Such structured information may include stock quotes, financial data, time tables, and customer records. (Spec. 1.) Presentation in the HTML format is convenient for human readers. Knowledge extraction therefrom for automated processing, he opines, is difficult because HTML formatted information contains irrelevant or repetitive explanatory text besides data of interest. (*Id.*)

The Appellant's invention parses several HTML-formatted web pages into individual tree structures and compares the trees using an isomorphism function. Based on specified criteria the invention extracts at least some systematic differences or similarities and outputs the extracted data in a desired format. (*Id.* 19.)

B. ILLUSTRATIVE CLAIM

Claim 1, which further illustrates the invention, follows.

1. A computer-implemented method of automatic data extraction from a plurality of HTML formatted input documents, comprising:

accessing a collection comprising a plurality of HTML formatted input documents;

automatically parsing the HTML formatting codes of each of the HTML formatted input documents into a

hierarchical tree structure having at least one root node and a sub-tree containing information data;

automatically storing the obtained document tree structure for each document in memory;

automatically performing a tree isomorphism function operation on each stored input document tree structure to compare the tree structures against each other and determine corresponding sub-trees based on the HTML formatting codes;

based on specified criteria, automatically extracting at least a subset of systematic differences and/or similarities obtained from a systematic comparison of information data contained within corresponding structural sub-trees of the plurality of HTML formatted input documents; and

automatically outputting extracted data in a desired target output format.

(Substitute Br.¹ A-1.)

III. EXAMINER'S REJECTION

Claims 1-30 stand rejected under 35 U.S.C. § 103(a) as obvious over U.S. Patent Application Pub. No. 2004/0199497 ("Timmons") in view of U.S. Patent No. 6,757,678 ("Myllymaki"). "Rather than reiterate the positions of parties *in toto*, we focus on an issue therebetween." *Ex parte Kuruoglu*, No. 2007-0666, 2007 WL 2745820, at *2 (BPAI 2007). The Examiner asserts, "Timmons teaches a method of automatic data extraction from multiple HTML pages, i.e., a collection of input documents, in which each source and target document is parsed into a tree structure (p. 9,

¹ We rely on and refer to the Substitute Brief on Appeal, in lieu of the original Brief on Appeal, because the latter was defective. We will not consider the original in deciding this appeal.

par. 129). The page is parsed as each container object is parsed one at a time; the container represents a sub-tree containing information data."
(Corrected Ans.² 4.) The Appellant argues that "paragraph [0129] of Timmons teaches parsing through individual container objects of a URL [i.e., uniform resource locator] and comparison of each individual object with an individual target tag. This is not a mapping and comparing of two document tree structures to each other since the target tag is a single object."
(Reply Br. 4.) Therefore, the issue is whether the Examiner has shown that Timmons parses the formatting codes of HTML documents into a hierarchical tree structure having a sub-tree.

"Both anticipation under § 102 and obviousness under § 103 are two-step inquiries. The first step in both analyses is a proper construction of the claims The second step in the analyses requires a comparison of the properly construed claim to the prior art." *Medichem, S.A. v. Rolabo, S.L.*, 353 F.3d 928, 933, (Fed.Cir. 2003) (internal citations omitted).

A. CLAIM CONSTRUCTION

"[T]he PTO gives claims their 'broadest reasonable interpretation.'" *In re Bigio*, 381 F.3d 1320, 1324 (Fed. Cir. 2004) (quoting *In re Hyatt*, 211 F.3d 1367, 1372 (Fed. Cir. 2000)). "Moreover, limitations are not to be read into the claims from the specification." *In re Van Geuns*, 988 F.2d

² We rely on and refer to the "corrected Examiner's Answer" (Supp. Reply Br. 1) in lieu of the original Examiner's Answer. We will not consider the original in deciding this appeal.

1181, 1184 (Fed. Cir. 1993) (citing *In re Zletz*, 893 F.2d 319, 321 (Fed. Cir. 1989)).

Here, claim 1 recites in pertinent part the following limitations: "automatically parsing the HTML formatting codes of each of the HTML formatted input documents into a hierarchical tree structure having at least one root node and a sub-tree containing information data . . ." Claim 17 includes similar limitations. Giving the two claims the broadest, reasonable construction, the limitations require parsing the formatting codes of HTML-formatted documents into a hierarchical tree structure having a sub-tree.

B. OBVIOUSNESS ANALYSIS

"In rejecting claims under 35 U.S.C. § 103, the examiner bears the initial burden of presenting a *prima facie* case of obviousness." *In re Rijckaert*, 9 F.3d 1531, 1532 (Fed. Cir. 1993) (citing *In re Oetiker*, 977 F.2d 1443, 1445 (Fed. Cir. 1992)). Here, the paragraph of Timmons relied on by the Examiner explains that "[t]he information retrieval processes of the present invention uses tags that have been generated previously to load a page of information and subsequently extract the desired information defined by the tag." (¶ [0129].) More specifically, the paragraph includes the following disclosure.

The URL of a page is passed with a "target" tag to the Feature Extraction indexer 500. The page is retrieved from the Web 502 and then each "container object" 504 is parsed one at a time. Each container is examined to see if "this container tag" equals the "target" tag 506. If this container matches the

target 508 then the information within this container is returned to the caller 510.

(*Id.*) In summary, the reference parses through individual container objects of a Web page and compares each individual object with an individual target tag. Contrary to the Examiner's assertion, we are unpersuaded that such a container object "represents a sub-tree . . ." (Corrected Ans. 4.) Therefore, we reverse the rejection of claims 1 and 17 and of claims 2-16 and 18-30, which depend therefrom.

IV. BOARD'S REJECTION

Under 37 C.F.R. § 41.50(b) (2007), we enter a new rejection against claim 1 under 35 U.S.C. § 103(a) as obvious over the combination of Myllymaki, the Appellant's admitted prior art ("AAPA"), and Timmons.

The question of obviousness is "based on underlying factual determinations including . . . what th[e] prior art teaches explicitly and inherently . . ." *In re Zurko*, 258 F.3d 1379, 1383 (Fed. Cir. 2001) (citing *Graham v. John Deere Co.*, 383 U.S. 1, 17-18 (1966); *In re Dembicza*k, 175 F.3d 994, 998 (Fed. Cir. 1999); *In re Napier*, 55 F.3d 610, 613 (Fed. Cir. 1995)). "'A *prima facie* case of obviousness is established when the teachings from the prior art itself would appear to have suggested the claimed subject matter to a person of ordinary skill in the art.'" *In re Bell*, 991 F.2d 781, 783 (Fed. Cir. 1993) (quoting *In re Rinehart*, 531 F.2d 1048, 1051 (CCPA 1976)).

Here, Myllymaki discloses "a method for merging tree data structures that contain redundant data, into more tractable tree data structures where those redundancies have been removed. Advantageously, Web users are able to retrieve information stored on one or more Web pages, available from one or more Web sites and locally merge the data." (Col. 2, ll. 29-35.) According to the reference "[a] group of . . . related input . . . documents . . . are obtained, for example, from various sources on the World Wide Web . . ." (Col. 5, ll. 61-63.) We find that obtaining these documents would have suggested the claim's "accessing a collection comprising a plurality of . . . formatted input documents . . ."

Myllymaki includes the following teachings about input documents.

Input document A can be represented by a data tree 140, that contains a root node, System A, designated by the reference numeral 155, with components CPU 160 that refers to the speed of the central processor unit, and Disk 165 that refers to the storage capacity, as well as CPU Option 170 and Component Disk Option 175. Similarly, input document B can be represented by a data tree 145, that contains a root node, System A, designated by the reference numeral 180, with components CPU 185 and Disk 190, as well as CPU Option 195 and Component Disk Option 200. Finally, input document C can be represented by a data tree 150, that contains a root node, System A, designated by the reference numeral 205, with components CPU 210 and Disk 215, as well as CPU Option 220 and Component Disk Option 225.

(Col. 7, ll. 22-35.) We find that representing input documents by such data trees would have suggested the claim's "automatically parsing the HTML formatting codes of each of the HTML formatted input documents into a

hierarchical tree structure having at least one root node and a sub-tree containing information data"

The Examiner finds, "Myllymaki teaches a method of mining electronic tree structured data, and automatically storing the tree structure for each document in memory (Col. 7, l[1]. 22-43)." (Corrected Ans. 4.) His finding is uncontested, and we, in turn, find that such storing would have suggested the claim's "automatically storing the obtained document tree structure for each document in memory"

In Myllymaki a "series of recursive passes . . . uncovers the fact that components 160, 185, and 210 refer to the same subcomponent, namely the CPU. As a result, similar sub-nodes 165, 190, and 215 that correspond to these nodes 160, 185, and 210 are determined . . . to be redundant." (Col. 7, ll. 50-56.) We find that uncovering such redundancy would have suggested the claim's "automatically performing a tree isomorphism function operation on each stored input document tree structure to compare the tree structures against each other and determine corresponding sub-trees based on the HTML formatting codes"

The reference's aforementioned "redundancies are removed in a MERGE process, resulting in an output data tree 260 of FIG. 8. In particular, two of three 'CPU' nodes and two of three 'Disk' nodes are removed, leaving a data tree with no redundancies, i.e., no redundant nodes. In this particular case, the remaining node 'CPU' is 160, the node 'Disk'

contained in the tree is 165." (*Id.* ll. 57-62.) Removing the redundancies serves to extract the nodes that are not redundant. We find that such extraction would have suggested the claim's "based on specified criteria, automatically extracting at least a subset of systematic differences and/or similarities obtained from a systematic comparison of information data contained within corresponding structural sub-trees of the plurality of HTML formatted input documents"

Myllymaki's "end product may be the data tree 260 of FIG. 8, or, alternatively, the information, e.g., the price structure" (*Id.* ll. 63-64.) We find that generating such an end product would have suggested the claim's "automatically outputting extracted data in a desired target output format."

Rather than operating on HTML-formatted documents as in the claim, however, Myllymaki discloses that "the specific embodiments of [its] invention that have been described" (col. 9, ll. 35-36) operate on input documents formatted in the eXtensible Markup Language ("XML"). (E.g., col. 5, ll. 29-30.) "[W]hen a patent 'simply arranges old elements with each performing the same function it had been known to perform' and yields no more than one would expect from such an arrangement, the combination is obvious." *KSR Int'l v. Teleflex Inc.*, 127 S. Ct. 1727, 1739 (2007) (quoting *Sakraida v. Ag Pro, Inc.*, 425 U.S. 273, 282 (1976)). More specifically, "when a patent claims a structure already known in the prior art that is altered by the mere substitution of one element for another known in the

field, the combination must do more than yield a predictable result." *Id.* at 1740 (2007) (citing *United States v. Adams*, 383 U.S. 39, 50-51 (1966)).

Here, the Appellant admits that HTML and its function were known in the art. More specifically, AAPA explains that "[s]tructured information is becoming increasingly present on the Internet in HTML format. Such structured information may include, for example, stock quotes, financial data, time tables, customer records, etc." (Spec. 1.) Not only does Myllymaki confirm that HTML and its function were known in the art, the reference describes similarities that were known between HTML and XML. To wit, both are standard languages used for web documents. (Col. 3, ll. 62-64, 66-67; col. 4, ll. 58-59.) During a document authoring stage, more specifically, an author embeds tags within the informational content of both an HTML document and of an XML document. (Col. 3, ll. 64-66; col. 4, ll. 60-61.) When a web server transmits the HTML document or the XML document to a web browser, the browser interprets the respective tags thereof and uses these tags to parse and display the document. (Col. 3, l. 66 to col. 4, l. 2; col. 4, ll. 61-65.) HTML tags and XML tags both can be used to create hyperlinks to other web documents. (Co. 4, ll. 3-5, 65-67.) Not only does Timmons confirm that HTML and its function were known in the art, moreover, "Timmons, p. 1, par. 9; states that web pages can be defined in both HTML and XML" as found by the Examiner. (Corrected Ans. 16.)

Based on the aforementioned findings, we conclude that the claim would have been obvious because the substitution of HTML-formatting for

XML-formatting would have yielded predictable results to one of ordinary skill in the art at the time of the invention. Therefore, we reject claim 1 as being obvious over the combination of Myllymaki, Timmons, and AAPA.

In an *ex parte* appeal, the Board of Patent Appeals and Interferences "is basically a board of review -- we review . . . rejections made by patent examiners." *Ex parte Gambogi*, 62 USPQ2d 1209, 1211 (BPAI 2001). Accordingly, we leave any further determination of the obviousness of claims 2-30 in view of Myllymaki, AAPA, and Timmons to the Examiner.

VI. ORDER

In summary, the rejection of claims 1-30 under § 103(a) is reversed. A new rejection under § 103(a), moreover, is entered against claim 1.

37 C.F.R. § 41.50(b) provides that "[a] new grounds of rejection pursuant to this paragraph shall not be considered final for judicial review." Section 41.50(b) also provides that, within two months from the date of the decision, the appellant must exercise one of the following options to avoid termination of proceedings of the rejected claims:

(1) Reopen prosecution. Submit an appropriate amendment of the claims so rejected or new evidence relating to the claims so rejected, or both, and have the matter reconsidered by the examiner, in which event the proceeding will be remanded to the examiner. . . .

(2) Request rehearing. Request that the proceeding be reheard under 37 C.F.R. § 41.52 by the Board upon the same record. . . .

Appeal 2007-2627
Application 10/248,681

No time for taking any action connected with this appeal may be extended under 37 C.F.R. § 1.136(a)(1)(iv)(2007).

REVERSED

37 C.F.R. § 41.50(b)

rwk

OLIFF & BERRIDGE, PLC.
P.O. BOX 320850
ALEXANDRIA VA 22320-4850