E

WIPO

SCIT/SDWG/8/4
ORIGINAL: English
DATE: February 23, 2007

WORLD INTELLECTUAL PROPERTY ORGANIZATION
GENEVA

STANDING COMMITTEE ON INFORMATION TECHNOLOGIES

STANDARDS AND DOCUMENTATION WORKING GROUP

Eighth Session
Geneva, March 19 to 22, 2007

REVISION OF WIPO STANDARD ST.22 (TASK No. 37)

Document prepared by the Secretariat

1. The Standards and Documentation Working Group (SDWG) of the Standing Committee on Information Technologies (SCIT), at its seventh session in May/June 2006, agreed to create a task for the revision of WIPO Standard ST.22 (Task No. 37), as had been proposed by the International Bureau and the European Patent Office. The SDWG established the ST.22 Task Force to handle such revision and appointed the International Bureau (IB) as Task Force Leader. The SDWG also requested the Task Force:

(a) to examine the use of non-Latin based characters, so that the Standard could cover non-Latin character based languages at a later stage;

(b) to examine the optical character recognition (OCR) accuracy rate (currently better than 98.5%) to see if a higher accuracy rate should be specified;

(c) to take into consideration what font styles and font sizes should be recommended for optimizing readability, for screen presentation, and OCR.
(See document SCIT/SDWG/7/9, paragraphs 32 to 38.)

2. In accordance with the above-mentioned decision by the SDWG, the IB issued circular C.SCIT 2628, dated June 23, 2006, inviting those Offices wishing to participate actively in the discussions to nominate a representative to the ST.22 Task Force.

3. Following the set-up of the corresponding electronic forum, the Task Force started its discussions, in July 2006, on the basis of an initial proposal presented by the Leader of the Task Force, with a view to making a final proposal concerning the revision of WIPO Standard ST.22 referred to in paragraph 1, above. The Task Force discussed the initial proposal, and, subsequently, made some revised proposals that took into consideration the comments provided by its members. The results of the said discussions are contained in the proposed revision of the Standard that is reproduced in the Annex to this document for consideration and approval by the SDWG. An oral report by the Task Force Leader will also be presented at the eighth session of the SDWG.

4. It should be noted that, at its seventh session, the SDWG agreed the idea that the revised Standard ST.22 should be brought to the attention of the applicants, offices, commercial information providers and patent attorneys. (See document SCIT/SDWG/7/9, paragraph 39.)

5. The SDWG is invited:

(a) to note the oral report of the Leader of the ST.22 Task Force referred to in paragraph 3, above;

(b) to consider and adopt the proposal concerning the revision of WIPO Standard ST.22 reproduced in the Annex to this document; and

(c) to consider, if deemed appropriate, requesting the IB to issue a circular to inform industrial property offices (IPOs) and SDWG Members of the revision of WIPO Standard ST.22 and to invite IPOs to publish a notice on their patent gazettes and Websites in order to brought that revised Standard to the attention of the applicants, commercial information providers and patent attorneys of their respective geographical areas, as referred to in paragraph 4, above.

[Annex follows]

ANNEX

STANDARD ST.22

RECOMMENDATION FOR THE AUTHORING OF PATENT APPLICATIONS FOR THE PURPOSE OF FACILITATING OPTICAL CHARACTER RECOGNITION (OCR)

Revision proposed by the ST.22 Task Force

INTRODUCTION

1. This recommendation applies to patent applications submitted on paper or submitted electronically (e-filed) but having the text body of the application submitted in image form (e.g., PDF or TIFF images).

2. This Recommendation has been established so as to assist in the preparation of a patent application in a typewritten form suitable for the subsequent production of an electronic digitized record of the contents of the patent application by the use of Optical Character Recognition (OCR) equipment.

3. This Recommendation has been established based upon the experiences of various offices in the use of OCR equipment. It has been drawn up with the objective of achieving the lowest possible error rate in the step of automatic reading of the text of patent applications whilst, at the same time, still permitting efficient personal reading of the document.

4. The primary aim of producing a digitized record of a patent application is to permit the easy publication of that application in a composed format using computer typesetting techniques and to thus enhance the presentation and value of patent documents to the advantage of all users. A further aim is to create a machine-readable database of the full text of a published document so that advantage can be taken at a later date of the possibilities offered by full text computer search.

DEFINITION

5. For the purposes of this Recommendation, the expression “patent application” means applications for patents for invention, inventor’s certificates, utility certificates, utility models, patents or certificates of addition, inventor’s certificates of addition and utility certificates of addition.

6. A mathematical or chemical formula is said to be “complex” if it cannot be displayed as a linear sequence of characters, each character having an optional subscript or superscript attribute. A formula is notably complex if it contains nested subscript/superscripts or if it contains the sum, integral or product mathematical symbols.

CREATION OF THE ORIGINAL

7. A patent application will often be prepared using word processing equipment. Experience has shown that the most efficient format of type that is to be used which would enable OCR equipment to be reliable is that defined in the International Standard Organization (ISO) Standard 1073/II, the so-called OCR-B format.

PAPER SUPPORT IF FILED ON PAPER

8. To facilitate scanning, the paper support of the typed application should have the following characteristics:

(a) the paper should be strong, white and substantially free of wood cellulose;

(b) the paper weight should lie between 70 and 120 gms/m^2;

(c) the paper size should preferably be A4, viz. 210 mm x 297 mm or 8 1/2 by 11 inches (which is the de-facto North-American standard);

(d) sheets should be free of creases, holes and should not be rolled;

(e) the paper should not be absorbent in order to avoid migration of the ink (for example, when using an ink jet printer).

PAGE LAYOUT RECOMMENDATIONS

9. The characters should be solid black on a white background.

10. A minimum margin of 2 cm should be present at the top, bottom and sides of each sheet, and one of 2.5 cm on the left side of each sheet. Any applicant's or representative's references should appear in the margin at the top. Please refer to Appendix 1.

11. Line numbering should be avoided. If required, line numbers should be typed using Arabic characters in the left hand margin area, at least 1 cm outside of the box as shown in Appendix 1. The font size of the line numbers should at least be 12 points.

12. Page numbering should be given by simply using Arabic characters without other delimiting characters. Page numbers should preferably be centered at the top or bottom of the sheet in the margin, as shown in Appendix 1.

13. The description, the claims and the abstract should be typed starting each on a new page. Moreover, the first word printed on the first page of each of the three aforementioned parts of the application should specify the corresponding part (in the language of the application); claims paragraphs should be numbered sequentially. The preferred format for the numbering of the claims is to use decimal numbers followed by a point and a tabulation.

14. Pages should be constituted of single column paragraphs (text paragraphs or paragraphs containing an embedded image).

15. Pages containing paragraphs must have a portrait orientation.

16. Landscape orientation should be avoided. It is acceptable only for pages containing embedded drawings or tables that would not fit in a portrait orientation.

17. Any page contains only one direction of text.

18. Landscape pages should be turned 90 degrees counterclockwise for integration within the set of portrait pages.

19. The use of footnotes, footers, margin texts and headers must be avoided (except page numbering).

PARAGRAPH LAYOUT RECOMMENDATIONS

20. A line of text should not contain tables, complex chemical or mathematical formulae.

21. Images and drawings should at maximum be included in the "Drawings" section of the patent application. If embedded images are required to assist the presentation and improve the understanding of the "description" and "claims" sections of the patent application, they should be easily separable from pure text paragraphs. It is advised that such items be separated from the text line above and below them by blank margins with a minimum height of 1 cm that run the entire width of the page; all text at the same horizontal location as a detected embedded image is likely not to be recognized as text and to be considered as part of the image.

22. Tables should be easily separable from pure text paragraphs. It is advised that tables be separated from the text line above and below them by blank margins with a minimum height of 1 cm that run the entire width of the page; paragraphs of texts must not be horizontally adjacent to tables.

23. Handwritten text paragraphs or annotations must be avoided. If required, they would be considered as embedded drawings and should follow the recommendation given in paragraph 21.

24. Typing should be done at one and a half line spacing.

25. Paragraphs should be separated by spacing that is at least twice as big as the intra-paragraph line spacing.

26. All characters within a paragraph line should have their baselines carefully aligned.

27. Justified text paragraphs must be avoided. If applied, the spacing between words should be at least as big as with unjustified text. Justified text may prevent the OCR systems to correctly identify the word boundaries in a paragraph.

28. Word splitting by the use of hyphens should be avoided (for example, at the end of lines or table cells).

TABLE RECOMMENDATIONS

29. Only white background should be used.

30. Tables must have borders. The borders should be thicker than 1.5 points and be only solid lines.

FONT RECOMMENDATIONS

31. The minimal recommended font size is 12 points, 14 points being preferred. As a general rule, all characters of a paragraph should have the same font size.

32. Text paragraphs containing subscripts and superscripts should use a font size of at least 12 points and recommended 14 points (the bigger, the better). It should be ensured that the bounding box of subscript or superscript characters intersects sufficiently with the bounding box of the normal characters on the same line (this prevents the OCR procedures to put the subscripts/superscripts on different lines).

33. The recommended fonts are the following in order:

(a) Monospaced family: OCR-B, Courier New

(b) Serif family: ITC Officina Serif, Times New Roman

(c) Sans Serif family: Verdana, ITC Officina Sans, Arial

However, the Arial and Times New Roman fonts are not recommended for applications containing chemical and/or mathematical formulae, as well as acronyms mixing letters and digits. For Chinese characters, the Song font is recommended.

34. The characters of the fonts have to be well shaped, with no shadows. The spaces between characters should be large enough (narrow spacing should be avoided).

35. Unusual characters should be avoided at maximum. If necessary, they belong preferably to the standard Greek alphabet and to the symbol font (by order of preference). Characters that cannot be found in the UNICODE range must not be used; those characters are recognized as embedded images by OCR engines and therefore make the recognized text difficult to read. Each office shall define and publish the character set it enforces for the preparation of the applications.

36. Narrow and cursive fonts should not be used.

37. Text should not be underlined. If required, it should be assured that the underline does not intersect with the underlined characters' bounding boxes.

38. Bold and italic styles should be avoided as much as possible.

RECOMMENDATIONS FOR NON-LATIN LANGUAGES

39. Within sections/pages of patent applications, the mixing of Latin and non-Latin languages is problematic for the OCR procedures and should be avoided.

CORRECTIONS

40. Corrections of the text of an application should be done by reprinting the whole page. Proof correction marks, as, for example, specified in the International Standard ISO 5776, are not accepted. Correction means such as white correcting fluid, self adhesive strips of paper, erasure or strikethrough are not accepted. Replacement pages shall not be sent by fax to the office using the 200 dpi resolution; pages should be sent physically or as an email attachment.

RECOMMENDATIONS FOR OFFICES

41. Patent offices should avoid altering the received pages before submitting them to scanning and OCR operations. For example, some current practices include stamping operations that may superimpose characters on pages, making text submitted by the applicant unreadable by OCR procedures. If stamps/changes have to be applied on the original pages, the office should take measures to ensure that the changes only occur in the margins of the documents, as defined in Appendix 1.

IMPLEMENTATION

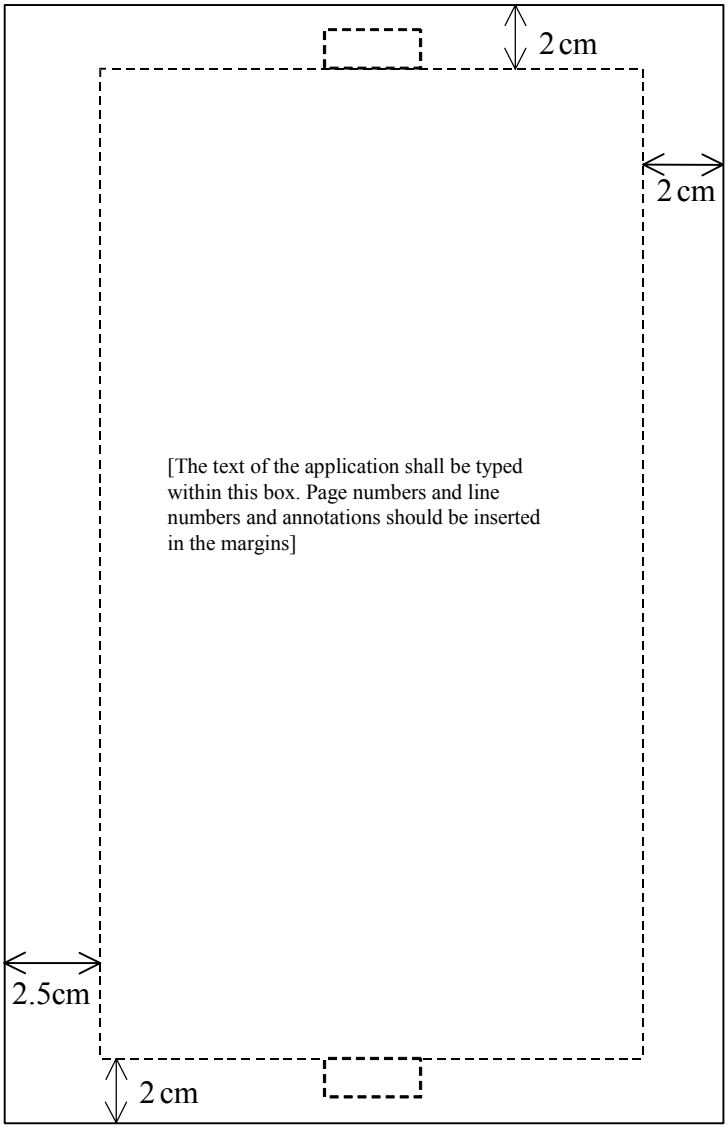
42. It is recommended that offices intending to start accepting or requesting the filing of patent applications typed in OCR format should publish full guidance in their Official Gazettes at regular intervals and on their websites, defining therein the exact character type or types permitted, and specifying the exact paper size allowable.

Examples

43. Examples of good and bad practices regarding OCR are reproduced in Appendix 2 to this Recommendation. The examples show what should and what should not be done, along with a short explanation.

[Appendices follow]

APPENDIX 1



Original Size = A4

[Appendix 2 follows]

APPENDIX 2

EXAMPLES OF GOOD AND BAD PRACTICES

This Appendix contains good and bad examples of patent document pages with respect to the accuracy obtained when performing OCR operations on them.

EXAMPLES OF GOOD PRACTICES

1. A good description page

Projection exposure system, method for manufacturing a micro-structured structural member by the aid of such a projection exposure system and polarization-optical element adapted for use in such a system

The invention relates to a projection exposure system, in particular for micro-lithography. The invention further relates to a method for manufacturing a micro-structured component and a polarization-optical element for the extreme ultraviolet (EUV) region.

For highest possible precision of the optical image to be obtained in complicated optical instruments such as a projection exposure system, the influence of the polarization of the light must be considered or, respectively, the polarization must be influenced specifically. For example, in particular in case of great incidence angles, polarization effects occur in the mirror systems, which projection exposure systems in the EUV region are based on, for lack of suitable transparent materials. These polarization effects are in particular due to the varying reflectivity of the mirrors for s-polarized and p-polarized light and can give rise to imaging errors or other undesired effects. Efforts have been made to measure possible polarization effects in the individual components of projection exposure systems.

For example, EP 1 306 665 A2 discloses an optical instrument for measuring polarization-dependent properties which comprises a light source in the EUV or X-radiation region and a rotatable polarizer. The polarizer is substantially comprised of a set of mirrors that reflects the incident light at least three times. The mirrors are arranged in such a way that the optical axes of the incident and emergent light are on the same straight line.

2. A good claims page

Claims

1. A projection exposure system, in particular for micro-lithography, comprising a light source (18) for producing light in the EUV region; a first optical system (19, 20, 21, 22, 23, 24) for illuminating a mask (25) by the light of the light source (18); and a second optical system (26, 27, 28, 29, 30) for imaging the mask (25) on a structural member (32); wherein at least one polarization-optical element (1) for the EUV region is disposed on the beam path between the light source (18) and the structural member (32), said at least one polarization-optical element comprising at least one reflective cone surface (3, 7, 12, 14), having a polarizing effect for the light produced by said light source.

2. Projection exposure system according to claim 1, wherein the polarization-optical element comprises

- at least one cone element (2) having an outer cone surface (3) being reflective for at least a given polarization component of the EUV light being incident under an angle of inclination α to the axis of rotational symmetry of the cone surface in a region of inclination angles α between 0° and a maximum angle of inclination such that the cone surface (3) has a polarizing effect, and
- having at least one further reflective component (4; 11, 13) for the EUV light to bundle the EUV light being reflected from the outer cone surface (3) of the cone element (2).

3. Projection exposure system according to claim 1 or 2, wherein the polarization-optical element comprises

- at least one further cone element (6) having an outer cone surface (3) being reflective for at least a given polarization component of the EUV

All recommendations are met: margins, a standard font (Times New Roman), a good font size, line numbers big enough and separated from the text, no justification, limited use of bold, no italics, no underlined text....

3. A good complex description page

[0134] When performing the first iteration of step S9-4, the values of D_a, A_a, D_b and A_b are the values previously calculated at step S7-2, while all values of λ_n are zero.

[0135] The equations used by solver 244 at step S9-6 comprise the following in this embodiment:

$$if \left(\lambda^{n+1}_{x,y,z-\max}\right)_{ang\&lin} < 0 \; then \; \lambda^{n+1}_{x,y,z-\max} = 0 \quad (46)$$

$$if \left(\lambda^{n+1}_{x,y,z-\min}\right)_{ang\&lin} > 0 \; then \; \lambda^{n+1}_{x,y,z-\min} = 0 \quad (47)$$

$$\lambda^{n+1}_{lin} = \lambda^{n+1}_{lin-\min} + \lambda^{n+1}_{lin-\max} \quad (48)$$

$$\lambda^{n+1}_{ang} = \lambda^{n+1}_{ang-\min} + \lambda^{n+1}_{ang-\max} \quad (49)$$

[0136] The equations used by solver 244 at step S9-8 comprise the following in this embodiment:

$$D^{n+1}_a = D^n_a + L\frac{\left(\lambda^{n+1}_{lin} - \lambda^n_{lin}\right)}{m_a} \quad (50)$$

$$A^{n+1}_a = A^n_a + I^{-1}_a[r^s_a]L.\left(\lambda^{n+1}_{lin} - \lambda^n_{lin}\right) + I^{-1}_a T\left(\lambda^{n+1}_{ang} - \lambda^n_{ang}\right) \quad (51)$$

$$D^{n+1}_b = D^n_b - L\frac{\left(\lambda^{n+1}_{lin} - \lambda^n_{lin}\right)}{m_b} \quad (52)$$

$$A^{n+1}_b = A^n_b - I^{-1}_b[r^s_b]L.\left(\lambda^{n+1}_{lin} - \lambda^n_{lin}\right) - I^{-1}_b T\left(\lambda^{n+1}_{ang} - \lambda^n_{ang}\right) \quad (53)$$

[0137] Referring again to Figure 7, at step S7-6, solver 244 performs a convergence test. In this embodiment, solver 244 performs processing to determine whether the values of λ calculated for the current iteration differ from the values of λ calculated for the previous iteration by more than a predetermined threshold, in accordance with the following equation:

$$\sum_{\lambda} \frac{\left(\lambda^{n+1} - \lambda^n\right)^2}{\lambda^{n2}} \leq Threshold \quad (54)$$

[0138] In this embodiment, the threshold employed in Equation (54) is set to 10^{-4}.

[0139] At step S7-8, solver 244 determines whether a predetermined number of iterations of the processing at steps S7-2 to S7-8 have been performed. In this embodiment, solver 244 determines whether 50 iterations have been performed.

The text paragraphs do not contain unusual mathematical characters. The mathematical formulae are correctly embedded allowing for an easy segmentation of the embedded images by the OCR engines. A possible result of the segmentation is shown in blue.

EXAMPLES OF BAD PRACTICES

1. A poor quality page with many defects

LOGICAL VIEW REPORT

Logical View

This is the high-level architecture for the POST Tools system. The logical view of the high-level architecture is the conceptual view, employing the SEI ABD decomposition representation of design elements, implementations, templates, requirements, decisions, and constraints.

Architecture

The Architecture package consists of the functional subsystem decomposition in design elements and implementations, the templates corresponding to each subsystem, and the deployment and concurrency views of the architecture.

Subsystems

The conceptual architecture subsystems are design elements and design element branches representing successive decompositions in a tree. These design elements are decomposed by functionality, quality scenarios, and constraints.

Browser Services

The user's desktop PC browser implements the save, print and view report functionalities. This includes the capabilities provided by browser plug-ins.

Database Query

An SQL database query and result set, using Java and JDBC/ODBC.

HyperText Transfer Protocol

The system employs a standard HTTP interface for client-server communication.

Java Applets

This is a generic implementation class representing a Java applet. In our design the applet probably is implemented within the Silver Stream application server development and run-time environment.

Java Beans

This implementation suggests any Java component in a plug-in style component framework, such as a Java Bean or an Enterprise Java Bean.

Java Servlets

This implementation suggests any Java Servlet managed by an application server or web server. Java Server Pages implementations might also be suitable alternatives.

POST Tools

This class represents the top-level design element in our architecture, representing all functionality from Vision and stakeholder input, legacy system constraints, inter-project operation and dependencies, and other contributing factors. We decompose all ABD functionality from this design element.

Table Lookup

This is a generic implementation class suggesting that the system performs a table lookup to find a component given an index key.

Paragraph 9 is not respected (the page was probably submitted by fax at 200 dpi to the office –please note the “noise”– and some text appears on heavy gray backgrounds). Paragraphs 10 and 41 are not respected; a reference number (604115530.100802) is stamped within the body of the page (it should be in the margins). The page numbering is incorrect (should be 9, not “page 9” see paragraph 12). Last but not least, the font size is too small (paragraph 31). Such pages should ideally not be accepted by offices and replacement pages requested (this page is impossible to OCR correctly).

2. A page with a non-white background

REVENDICATIONS

1. Dispositif d'usinage (D) du type de celui associant une machine-outil d'usinage (100) à un dispositif porte-pièce (200) équipé d'un axe de mise en mouvement de rotation transversal (A) par rapport à l'axe de plongée (Z), CARACTÉRISÉ PAR LE FAIT QUE le dispositif porte-pièce (200) est constitué par un bâti (210) supportant deux paliers de guidage (210 et 230) en rotation selon ledit axe de rotation transversal (A), la structure formée par le bâti (210) et les deux paliers (220 et 230) étant fermée par la pièce à usiner (300) dont les extrémités viennent se fixer auxdits paliers (310 et 220), la pièce à usiner (300) étant une pièce longue du type de celle comportant des surfaces à usiner concentrées à ses deux extrémités ET PAR LE FAIT QUE la machine-outil (100) est du type de celle assurant la mise en mouvement de deux coulants porte-outil indépendants (110 et 120) de façon à ce que les usinages des deux extrémités de la pièce (300) soient réalisés par un coulant différent.

2. Dispositif d'usinage (D) selon la revendication 1, CARACTÉRISÉ PAR LE FAIT QUE chaque palier (220 et 230) comprend et guide un plateau tournant (221 et 231) équipé d'un moyen de mise en mouvement motorisé, la rotation des deux plateaux (221 et 231) étant synchronisée.

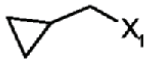
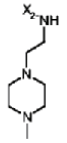
3. Dispositif (D) selon la revendication 2, CARACTÉRISÉ PAR LE FAIT QUE chaque plateau (220 et 230) est équipé de deux appuis (410, 420 et 510, 520) pour accueillir et maintenir en position l'extrémité de la pièce (300).

4. Dispositif (D) selon la revendication 1, CARACTÉRISÉ PAR LE FAIT QUE le bâti (210) du dispositif porte-pièce (200) est lui-même monté mobile en rotation selon un axe (B) perpendiculaire à l'axe (A) de rotation défini par les deux paliers (220 et 230) qu'il supporte.

5. Procédé d'usinage d'une pièce longue (300) du type de celle comportant des surfaces à usiner concentrées à

Paragraph 9 is not respected. The page needs to be filtered to attempt to remove the noisy background before submitting it to an OCR operation. If OCRed as is, the obtained text is unreadable.

3. A page with faint characters

#	R2	A	UV max [nm]:	MS (ESI) $(M+H)^+$	
25			305, 350	476	Trihyc 1,41 ((m, 2H (m, 1H

Beispiele 26-40

Die folgenden Verbindungen sind über ein analog beschrieben, hergestellt. Die Herstellung des Benz beschrieben. Das für die Darstellung des Amids ei

A small area of the page is zoomed to show the characters; the color of the original text is probably gray, resulting after the scanning in 300dpi black and white in characters which are not solid. As a result, the accuracy of the OCRed text is poor (Paragraph 9 is not respected).

4. A page with handwritten text

TITLED : JIG HEAD SWAY BAR

BACK GROUND

IN THE ART OF FISHING THERE IS A PIECE OF TACKLE KNOWN AS A PIVOT - HEAD JIG WHICH USES SPECIALIZED OR SPECIFICALLY SHAPED HOOKS TO PROVIDE AN ACTION PRODUCING LURE COMBINATION. MY INVENTION THE SWAY BAR ALLEVIATES THIS NEED FOR SPECIAL HOOKS BY BEING ABLE TO BOTH SUPPORT THE JIG HEAD AND ALLOW FOR CONNECTION OF OTHER REQUIRED TACKLE

As to be expected, the text obtained by OCRing this page is unreadable. Offices should request typewritten text to ensure minimum publication quality.

5. A page with a non recommended layout and other defects

relation to the determination of AN by FTIR spectroscopy

This concept is illustrated in Figure 1 for AN, the BN analysis being analogous but using a different reagent. Differential spectroscopy is then used to eliminate the spectral contributions from the base oil and any additives and/or contaminants and breakdown products present in the oil that may spectrally interfere with the measurement of the signal from the reaction product. This is achieved by treating a portion of the sample with a blank reagent, this portion effectively serving as a reference oil. Figure 2 illustrates the general analytical protocol.

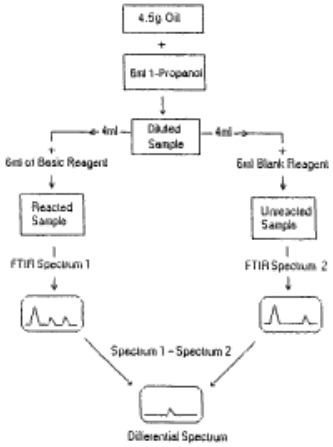


Figure 2. Analytical protocol for the determination of AN by FTIR spectroscopy.

In this procedure, the sample is first diluted with an innocuous solvent (1-propanol), then split and treated with a reactive and a blank reagent to produce two samples for spectral analysis. Since these two samples are the same except for the reaction products, subtraction of their spectra leaves only the spectral contribution related to AN.

The COAT AN/BN Analyzer

The COAT AN/BN Analyzer has been designed and programmed to automate AN/BN analyses based on the concepts laid out above. Figure 3 illustrates key components of the COAT AN/BN Analyzer: an FTIR spectrometer, a sample handling accessory, an autosampler, and the computer that controls the system.

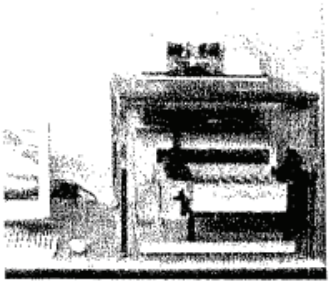


Figure 3. The COAT AN/BN Analyzer and its key components

The compact nature of the sample handling system is made possible by the dilution of the sample in the analytical protocol (Figure 2), allowing a micropump to be substituted for the peristaltic pump employed in most FTIR used oil analyzers. The resulting low viscosity of the sample dramatically

This page does not respect paragraph 14 (single column formatting), uses italic and bold fonts (against paragraph 38), has manual corrections performed after printing (against paragraph 40). The left-right justification of the paragraph is also not recommended (paragraph 27), although in this case, this would not have negative effects on the OCR since the words are still sufficiently separated by white spaces. Paragraph 24 is also not respected (one and a half line spacing).

6. A page with line numbers that are too small

[0028] Figs. 9A-9B are plots showing the percent of mitomycin C released from liposomes comprised of HSPC/mPEG-DSPE/lipid-DTB-mitomycin C (Fig. 9A) and HSPC/cholesterol/mPEG-DSPE/lipid-DTB-mitomycin C (Fig. 9B) as a function of time of incubation in the presence of cysteine at concentrations of 150 µM (closed symbols) and at 1.5 mM (open symbols);

[0029] Fig. 10 is a plot of growth rate of M109 cells, expressed as a percentage based on growth of M109 cells in the absence of drug and cysteine, as a function of mitomycin C amount, in nM, for free mitomycin c (open triangles), liposomes comprised of HSPC/mPEG-DSPE/lipid-DTB-mitomycin C (closed squares), and liposomes comprised of HSPC/cholesterol/mPEG-DSPE/lipid-DTB-mitomycin C (open circles);

[0030] Fig. 11A is a plot of growth rate of M109 cells, expressed as a percentage based on growth of M109 cells in the absence of drug or cysteine, as a function of mitomycin C concentration in nM. Shown are cells treated mitomycin C in free form (open triangles) and with mitomycin C in free form plus 1000 µM cystein (closed triangles). Also shown are cells treated with the liposome formulation comprised of HSPC/PEG-DSPE/lipid-DTB-mitomycin C (open circles) and with the liposome formulation with additional cysteine added at concentrations of 150 µM (open diamonds), 500 µM (closed circles) and 1000 µM (open squares);

[0031] Fig. 11B is a plot of growth rate of M109 cells, expressed as a percentage based on growth of M109 cells in the absence of drug or cysteine, as a function of mitomycin C concentration in nM. Shown are cells treated mitomycin C in free form (open triangles) and with mitomycin C in free form plus 1000 µM cysteine (closed triangles). Also shown are cells treated with the liposome formulation comprised of HSPC/cholesterol/mPEG-DSPE/lipid-DTB-mitomycin C (open circles) and with the liposome formulation with additional cysteine added at concentrations of 150 µM (open diamonds), 500 µM (closed circles) and 1000 µM (open squares);

[0032] Fig. 12 is a plot showing the percent increase in cytotoxicity (as determined by ($IC50_{no\ cysteine}$/$IC50_{cysteine}$)x100)) of free mitomycin C (closed squares), mitomycin C associated with liposomes comprised of HSPC/cholesterol/mPEG-DSPE/lipid-DTB-mitomycin C (closed circles), and liposomes comprised of HSPC/mPEG-DSPE/lipid-DTB-mitomycin C (open triangles) to M109 cells *in vitro* at various concentrations of cysteine;

[0033] Fig. 13A is a plot showing the concentration of mitomycin C in the blood of

---

Line numbers cause problems to the OCR engines for several reasons (paragraph 11):

- They may not be aligned with the lines they correspond to, leading to baseline detection defaults.
- They could be too small, leading to recognition errors that would prevent the XML extraction procedures to remove them correctly from the text body of the page.
- They could be misplaced within the body text area of the page, or in the margins but too close to the body text area and as a result will appear inside the text stream exported by the OCR operations.

In this example, they are too small.
Subscript characters are also too small in this example (paragraph 32).

7. A page containing several directions of text

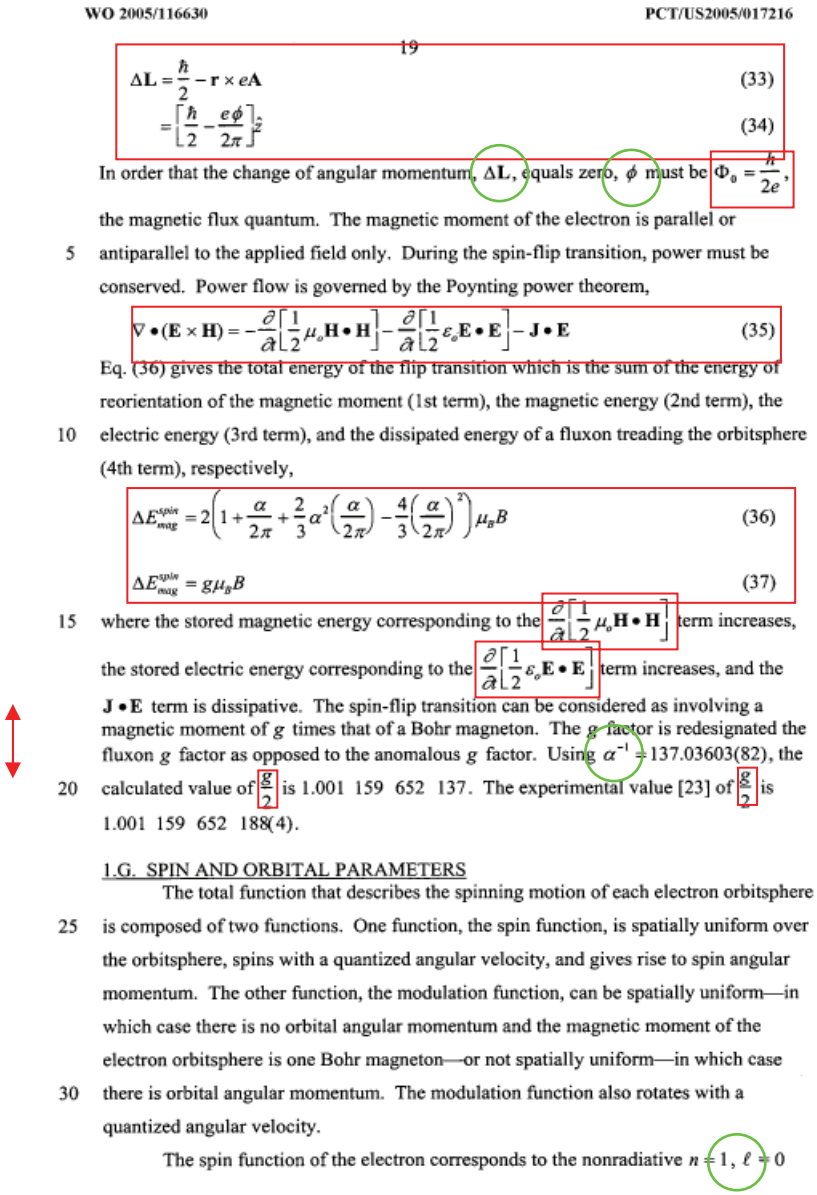
Table 11 (continued-2)

		Saponification degree	Amount in retardation-controlling agent solution (mass parts)					Amount in UV absorber solution (mass parts)						Mixing ratio of solutions			
			Retardation-control agent A-2	Retardation-control agent A-12	Retardation-control agent B	Retardation-control agent C	Retardation-control agent D	UV absorber A	UV absorber B	UV absorber C	UV absorber D	UV absorber E	UV absorber F	Cellulose acetate solution	Matting agent solution	Retardation-controlling agent solution	UV absorber solution
This invention	Cellulose acetate film 25			6	14								15	94.6	1.2	6.2	1.3
This invention	Cellulose acetate film 26		3	3	14								15	94.6	1.2	7.0	3.2
This invention	Cellulose acetate film 27		5	5	10								15	94.6	1.2	6.2	0.8
This invention	Cellulose acetate film 28	59.8		5			15	4.8	10.2					94.6	1.2	6.2	0.8
This invention	Cellulose acetate film 29			10			10	4.8	10.2					94.6	1.2	6.2	0.8
This invention	Cellulose acetate film 30						15	4.8	10.2					94.6	1.2	6.2	0.8
Comparative example	Cellulose acetate film 31	60.9	10	10										94.6	1.2	6.6	0
Comparative example	Cellulose acetate film 32	60.5			20									94.6	1.2	4.1	0
Comparative example	Cellulose acetate film 33	60.9									5	10		94.6	1.2	0	6.3
Comparative example	Cellulose acetate film 34	60.8	10	10				10.5	4.5					94.6	1.2	7.1	0.8
Comparative example	Cellulose acetate film 35	60.8	10	10				10.5	4.5					94.6	1.2	7.1	0.8

This example does not respect paragraph 17.

One of the limitation of the best OCR engines available today is that they can read only one direction of text in one page (a preprocess of the page is to detect the main text orientation of the page). As a result, all the words that are not in the main text direction are ignored. It is of course acceptable to have in a page a landscape table or even a main landscape text with portrait annotations in the margins (page number, application number and so on).

8. A page with mixed embedded mathematical formulae and text



WO 2005/116630 PCT/US2005/017216

19

$$\Delta \mathbf{L} = \frac{\hbar}{2} - \mathbf{r} \times e\mathbf{A} \quad (33)$$

$$= \left[\frac{\hbar}{2} - \frac{e\phi}{2\pi}\right]\hat{z} \quad (34)$$

In order that the change of angular momentum, $\Delta \mathbf{L}$, equals zero, ϕ must be $\Phi_0 = \frac{\hbar}{2e}$, the magnetic flux quantum. The magnetic moment of the electron is parallel or antiparallel to the applied field only. During the spin-flip transition, power must be conserved. Power flow is governed by the Poynting power theorem,

$$\nabla \bullet (\mathbf{E} \times \mathbf{H}) = -\frac{\partial}{\partial t}\left[\frac{1}{2}\mu_o \mathbf{H} \bullet \mathbf{H}\right] - \frac{\partial}{\partial t}\left[\frac{1}{2}\varepsilon_o \mathbf{E} \bullet \mathbf{E}\right] - \mathbf{J} \bullet \mathbf{E} \quad (35)$$

Eq. (36) gives the total energy of the flip transition which is the sum of the energy of reorientation of the magnetic moment (1st term), the magnetic energy (2nd term), the electric energy (3rd term), and the dissipated energy of a fluxon treading the orbitsphere (4th term), respectively,

$$\Delta E_{mag}^{spin} = 2\left(1 + \frac{\alpha}{2\pi} + \frac{2}{3}\alpha^2\left(\frac{\alpha}{2\pi}\right) - \frac{4}{3}\left(\frac{\alpha}{2\pi}\right)^2\right)\mu_B B \quad (36)$$

$$\Delta E_{mag}^{spin} = g\mu_B B \quad (37)$$

where the stored magnetic energy corresponding to the $\frac{\partial}{\partial t}\left[\frac{1}{2}\mu_o \mathbf{H} \bullet \mathbf{H}\right]$ term increases, the stored electric energy corresponding to the $\frac{\partial}{\partial t}\left[\frac{1}{2}\varepsilon_o \mathbf{E} \bullet \mathbf{E}\right]$ term increases, and the $\mathbf{J} \bullet \mathbf{E}$ term is dissipative. The spin-flip transition can be considered as involving a magnetic moment of g times that of a Bohr magneton. The g factor is redesignated the fluxon g factor as opposed to the anomalous g factor. Using $\alpha^{-1} = 137.03603(82)$, the calculated value of $\frac{g}{2}$ is 1.001 159 652 137. The experimental value [23] of $\frac{g}{2}$ is 1.001 159 652 188(4).

1.G. SPIN AND ORBITAL PARAMETERS

The total function that describes the spinning motion of each electron orbitsphere is composed of two functions. One function, the spin function, is spatially uniform over the orbitsphere, spins with a quantized angular velocity, and gives rise to spin angular momentum. The other function, the modulation function, can be spatially uniform—in which case there is no orbital angular momentum and the magnetic moment of the electron orbitsphere is one Bohr magneton—or not spatially uniform—in which case there is orbital angular momentum. The modulation function also rotates with a quantized angular velocity.

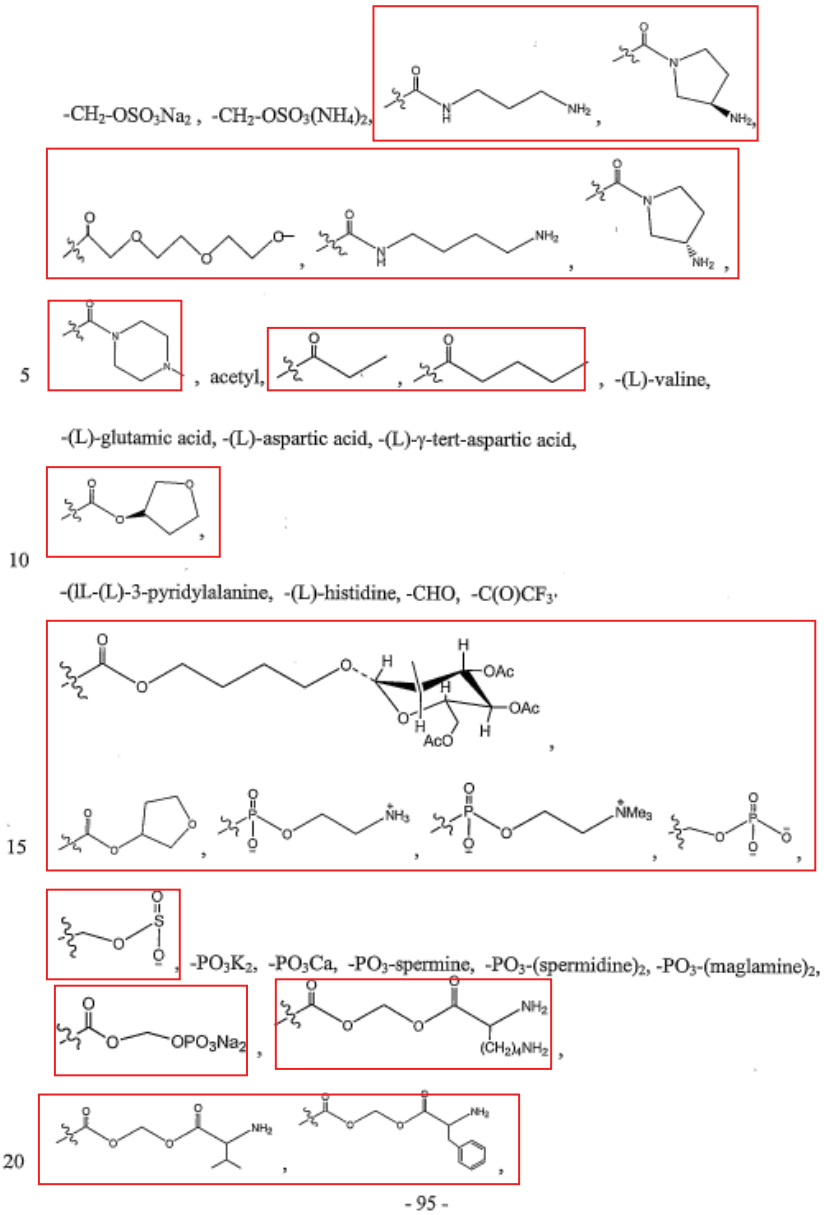
The spin function of the electron corresponds to the nonradiative $n = 1$, $\ell = 0$

This example does not respect paragraphs 20 and 22. The OCR engine is not able to separate correctly the text and the formulae (see the result of a manual segmentation of the formulae in red; the embedded formulae even intersect).

As a general comment, in this example, the text and the formulae are too dense for good recognition; paragraphs 24 and 25 are also not respected.

This example also uses unusual characters; Greek symbols can be used even if they increase the recognition difficulty of the page (see paragraph 35). However, it is highly unrecommended to combine italics, bold or underlined fonts with unusual characters (paragraph 38).

9. A page with mixed embedded chemical formulae and text



This example does not respect paragraphs 20 and 21. You can find in red one expected result of the drawings segmentation (done manually). This segmentation cannot be performed correctly by an OCR engine since the formulae are too close to the surrounding text.

10. A page with subscript characters that are too small

R_1 is hydrogen, C_1-C_6alkyl, C_2-C_6alkenyl, C_2-C_6alkynyl, C_1-C_6alkoxy, C_1-C_6haloalkyl, C_1-C_6haloalkoxy, (C_3-C_7cycloalkyl)C_0-C_4alkyl;

R_3 is selected from alkoxy, cycloalkoxy, phenyl, 4- to 7-membered heterocycles, -O$(CH_2)_m$phenyl, -O$(CH_2)_m$pyridyl, -E-$(CR_CR_D)_m$-Q, and Q, each of which is substituted with between 0 and 3 substituents selected from halogen, cyano, hydroxy, oxo, $(CR_AR_B)_j$-T, C_{1-6}alkyl, C_{1-6}alkoxy, C_{1-6}haloalkyl, C_{1-6}haloalkoxy, mono- and di-(C_{1-6}alkyl)amino, (C_{1-6}alkyl)(($CR_AR_B)_j$-T)amino, benzyl, S(O)$_i$(C_{1-6}alkyl), α,ω-C_{1-4}alkylene, α,ω-C_{1-4}alkyleneoxy, α,ω-C_{1-4}alkylenedioxy, -E-$(CH_2)_m$-Q, and Q;

T is CO_2H, $CONH_2$, C_{1-6}alkoxycarbonyl, mono- or di-(C_{1-6}alkyl)aminocarbonyl, SO_3H, SO_2NH_2 or SO_2(C_{1-6}alkyl);

j is an integer ranging from 0 to 6;

Q is a saturated heterocyclic ring comprising between 4 and 7 ring members, in which the point of attachment is a carbon or nitrogen atom;

E is O, NR_D, or a single covalent bond;

R_8 and R_9 are independently chosen from hydrogen, halogen, hydroxy, C_1-C_6alkyl, C_1-C_6alkenyl, (C_3-C_6cycloalkyl)C_0-C_4alkyl and C_1-C_6alkoxy; and

Ar is phenyl which is mono-, di-, or tri-substituted; or 1-naphthyl, 2-naphthyl, pyridyl, pyrimidinyl, pyrazinyl, pyridizinyl, thienyl, thiazolyl, pyrazolyl, imidazolyl, tetrazolyl, oxazolyl, isoxazolyl, pyrrolyl, furanyl, indolyl, indazolyl, or triazolyl, each of which is optionally mono-, di-, or tri-substituted.

Yet other compounds of Formula VIII include those compounds in which the group designated:

is chosen from naphthyl, tetrahydronaphthyl, benzofuranyl, benzodioxolyl, indanyl, indolyl, indazolyl, benzodioxolyl, benzo[1,4]dioxanyl and benzoxazolyl, each of which is substituted with from 0 to 3 substituents independently chosen from R_x.

Certain compounds of Formula IX include those in which

Ar is mono-, di-, or tri-substituted phenyl, which phenyl group is substituted with one to three substituents independently chosen from hydroxy, halogen, cyano, amino, nitro, -COOH, aminocarbonyl, $-SO_2NH_2$, C_{1-6}alkyl, C_{1-6}alkenyl, C_{1-6}alkynyl, C_{1-6}haloalkyl, C_{1-6}aminoalkyl, C_{1-6}hydroxyalkyl, C_{1-6}carboxyalkyl, C_{1-6}alkoxy, C_{1-6}haloalkoxy, C_{1-6}alkylthio, C_{1-6}alkanoyl, C_{1-6}alkanoyloxy, C_{1-6}alkanone, C_{1-6}alkyl ether, mono- or di-(C_{1-6}alkyl)aminoC_{0-6}alkyl, -NHC(=O)(C_{1-6}alkyl), -N(C_{1-6}alkyl)C(=O)(C_{1-6}alkyl), -NHS(O)$_n$(C_{1-6}alkyl), -(C_{1-6}alkyl)C(=O)NH_2, -(C_{1-6}alkyl)C(=O)NH(C_{1-6}alkyl), -(C_{1-6}alkyl)C(=O)NH(C_{1-6}alkyl)(C_{1-6}alkyl), -S(O)$_n$(C_{1-6}alkyl), -S(O)$_n$NH(C_{1-6}alkyl), -S(O)$_n$N(C_{1-6}alkyl)(C_{1-6}alkyl) and Z; or

---

This is a typical example where the subscripts characters are too small to allow for accurate recognition. This phenomenon is particularly recurrent for patents in the chemistry field.

11. A page with badly formatted tables

Table D

Other compounds of the invention result from selecting appropriate features from the table of possible features below. For example, compound A77 results from the following selections: none-morpholino-aryl-$OCH_2(CO)$-piperazine-CH_3.

Left-hand substituent	Left-hand ring	Aryl or heteroaryl	Ring substituent	Nitogen feature	Right-hand substituent
CH3	morpholino	aryl	OCH2	NHM	alkyl
isopropyl	piperazine	thiopene	OCH2(CO)	NMM	alkoxy
CH3CH2O(CO)CH2			SO2	morpholino	alcohol
none			OCH2(CO)OCH2	piperazine	substitued amine
				piperidine	acid
				pyrazole	ester
				pyrrolodine	$CH_2CH_2OCH_3$
					CH_2CH_2OH
					CH_2NH_2
					$CH_2NHCH_2CH_2CH_3$
					CH_2NHCH_3
					$CH_2NHCHCH_3CH_3$
					CH_3
					$CHCH_2CH_3$
					$COOCH_2CH_3$
					none

Table E

Other compounds of the invention result from selecting appropriate features from the table of possible features below. For example, compound B3 results from the following selections: none-morpholino-aryl-CH_2-piperazine-CH_2CH_2OH.

Left-hand substituent	Left-hand ring	Aryl or heteroaryl	Ring substituent	Nitogen feature	Right-hand substituent
CH3	morpholino	aryl	CH2	NHM	alkyl
isopropyl	piperazine	thiophene	CH2CH2	NMM	alkoxy
CH3CH2O(CO)CH2			CH2CH2CH2	morpholino	alcohol
none			CH2CH2CH2CH2	piperazine	substituted amine
				piperidine	acid
				pyrazole	ester
				pyrrolidine	$CH_2CH_2OCH_3$
					CH_2CH_2OH
					CH_2NH_2
					$CH_2NHCH_2CH_2CH_3$
					CH_2NHCH_3
					$CH_2NHCHCH_3CH_3$
					CH_3

In this example, the table boundaries are missing (against paragraph 30). As a result, the OCR engine will try to recognize contents of the tables as paragraph text. This leads to several other problems:

- The font size of the characters in the tables is too small (paragraphs 31 and 32).
- The baselines of the column headings are mixed (paragraph 26). As a result, the engine will wrongly detect subscripts or superscripts.
- The text stream obtained will not take into account the columns:

  Left-hand Left-Hand Aryl or Nitogen
  Substituent ring heteroaryl Ring substituent **feature** Right-hand substituent
  CH3…

12. A justified page

GKSS-Forschungszentrum Geesthacht GmbH, Max-Planck-Straße 1, 21502 Geesthacht

Verfahren zur Herstellung von Profilen aus Leichtmetallwerkstoff mittels Strangpressen

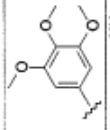
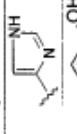
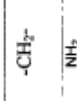
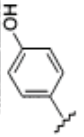
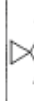
Beschreibung

Die Erfindung betrifft ein Verfahren zur Herstellung von Profilen aus Leichtmetallwerkstoff, insbesondere Magnesiumwerkstoff, mittels Strangpressen, bei dem ein Werkstoffvolumen durch eine Matrize, die die Form des gewünschten Profils bestimmt, zur Ausbildung des Profils gepreßt wird.

Die Herstellung von Profilen aus Leichtmetall- bzw. Leichtmetall-Legierungswerkstoffen mittels eines Strangpreßverfahrens ist eine allgemein eingeführte, bekannte Technologie und wird industriell angewendet. So ist es bekannt, daß konventionell verfügbare Leichtmetall- bzw. Leichtmetall-Knetlegierungen in Form von Gußblöcken durch konventionelles Strangpressen in Profilformen gepreßt werden. Dabei wird der Leichtmetall- bzw. Leichtmetall-Legierungsblock, im folgenden zusammenfassend kurz mit Werkstoffvolumen bezeichnet, bei Temperaturen

In this example, left and right justifications are applied to the paragraphs. If this makes the text look better, it sometimes makes OCR operations difficult when the separations between the words become too small (paragraph 27). This example also does not respect paragraph 28 that states that word splitting at the end of the lines should be avoided as much as possible (the OCR engine sometimes has difficulties distinguishing hard and soft hyphens, resulting in words containing undesired hyphens in the output).

13. A table with bad boundaries

Comp. No.	Exp. No.	Alk[a]	Y	Alk[b]	L	Physical data
106	B2	cb	C=O	$-CH_2-$		2R-trans
107	B3b	cb	C=O	$-CH_2-$		2R-trans
13	B8	cb	C=O			2R-trans, HCl(1:3); H_2O(1:1)
108	B2	cb	C=O			2R-trans HCl(1:2) H_2O(1:1)
109	B2	cb	C=O			2R-trans
110	B3b	cb	C=O			[2R-[2α,4β(E)]]
111	B2	cb	C=O			2R-trans

In this example, the boundaries of the table in the received original before scanning are of bad quality. After scanning, the OCR procedure is unable to detect correctly the table and a manual operation is required to segment the page. If such a page is not checked by an operator for quality, the text output will contain many undesired “junk” characters that will make the indexation of the document by search engines less effective.

14. Bad subscript and superscript characters

thiazolyl, pyrazolyl, pyridinyl, pyrimidinyl, purinyl, quinolinyl, benzofuran and isoquinolinyl.

p. "heteroaryl, optionally substituted," refers to a heteroaryl moiety as defined immediately above, in which up to 4 carbon atoms of the heteroaryl moiety may be substituted with a substituent, each substituent is independently selected from the group consisting of halogen, cyano, hydroxy, $(C_1\text{-}C_6)$alkyl, $(C_1\text{-}C_6)$alkoxy, $(C_1\text{-}C_2)$alkyl substituted with one or more halogens, $(C_1\text{-}C_2)$alkoxy substituted with one or more halogens, SR^8, and NR^8R^9, in which R^8 and R^9 are as defined above.

q. "heterocycle" or "heterocyclic ring" refers to any 3- or 4-membered ring containing a heteroatom selected from oxygen, nitrogen and sulfur; or a 5-, 6-, 7-, 8-, 9-, or 10- membered ring containing 1, 2, or 3 nitrogen atoms; 1 oxygen atom; 1 sulfur atom; 1 nitrogen and 1 sulfur atom; 1 nitrogen and 1 oxygen atom; 2 oxygen atoms in non-adjacent positions; 1 oxygen and 1 sulfur atom in non-adjacent positions; or 2 sulfur atoms in non-adjacent positions. The 5-membered ring has 0 to 1 double bonds, the 6- and 7-membered rings have 0 to 2 double bonds, and the 8, 9, or 10 membered rings may have 0, 1, 2, or 3 double bonds. The term "heterocyclic" also includes bicyclic groups in which any of the above heterocyclic rings is fused to a benzene ring, a cyclohexane or cyclopentane ring or another heterocyclic ring (for example, indolyl, quinolyl, isoquinolyl, tetrahydroquinolyl, benzofuryl, dihydrobenzofuryl or benzothienyl and the like). Heterocyclics include: pyrrolidinyl, tetrahydrofuranyl, tetrahydrothiophenyl, piperidinyl, piperazinyl, azepane, azocane, morpholinyl, isochroamyl and quinolinyl.

r. "heterocyclic, optionally substituted" refers to a heterocyclic moiety as defined immediately above, in which up to 4 carbon atoms of the heterocycle moiety may be substituted with a substituent, each substituent is independently selected from the group consisting of halogen, cyano, hydroxy, $(C_1\text{-}C_6)$alkyl, $(C_1\text{-}C_6)$alkoxy, $(C_1\text{-}C_2)$alkyl substituted with one or more halogens, $(C_1\text{-}C_2)$alkoxy substituted with one or more halogens, SR^8, and NR^8R^9, in which R^8 and R^9 are as defined above. Any nitrogen atom within such a heterocyclic ring

This example exhibits the following problems (paragraph 32):

- subscript and superscript characters too small;
- subscript characters located too low with respect to the baseline;
- superscript characters located too high with respect to the baseline.

As a result, lines 33 and 34 of the text are recognized as follows by the OCR procedure:

> "Substituted with one or more halogens, (C -C)alkoxy substituted
> 1 2
> 8 8 9 8 9
> with one or more halogens, SR , and NR R , in which R and R are"

15. An example with unusual characters

c = speed of sound in water;

$\tilde{z}_u$ = initial altitude for beam pair *u*;

$\Delta\varepsilon_{z,u} = \varepsilon_{z,p+1,u} - \varepsilon_{z,p,u}$ = comparable to sway-reduced altitude difference;

$\Delta\varepsilon_{\gamma,u} = \varepsilon_{\gamma,p+1,u} - \varepsilon_{\gamma,p,u}$ = comparable to sway-reduced horizontal displacem

$\varepsilon_{z,p,u}$ = difference of vertical linearization point in ping p, beam pair u, fror nominal $\tilde{z}_u$;

$\varepsilon_{z,p+1,u}$ = difference of vertical linearization point in ping p+1, beam pair u, nominal $\tilde{z}_u$;

$\varepsilon_{\gamma,p,u}$ = difference of horizontal-range sample v linearization point in ping p u, from the nominal $\gamma_{v,u}$. Note that this is the same for all horizonta samples;

$\varepsilon_{\gamma,p+1,u}$ = difference of horizontal-range sample v linearization point in ping pair u, from the nominal $\gamma_{v,u}$. Note that this is the same for all horiz samples;

$\gamma_{v,u}$ = nominal horizontal offset to horizontal-range sample u for beam pair

This sample exhibits the following problems:

- Unusual characters: Greek italic, and even characters with a tilda on top.
- The subscripts here again are too small.

With most OCR engines, all unusual characters will not be recognized correctly.

16. An example with narrow fonts and narrow spacing

23. The method of claim 18, wherein the data is encoded onto the representative transmission symbol by using a modulation method selected from a group consisting of: amplitude modulation, phase modulation, frequency modulation, single-sideband modulation, vestigial-sideband modulation, quadrature amplitude modulation, orthogonal frequency division modulation, pulse-code modulation, pulse-width modulation, pulse-amplitude modulation, pulse-position modulation, pulse-density modulation, frequency-shift keying, and phase-shift keying.

24. The method of claim 18, wherein each of the at least two communication signals is transmitted through a communication medium selected from a group consisting of: a wire medium, a wireless medium, an optical fiber ribbon, a fiber optic cable, a single mode fiber optic cable, a multi-mode fiber optic cable, a twisted pair wire, an unshielded twisted pair wire, a plenum wire, a PVC wire, and a coaxial cable.

25. The method of claim 18, wherein the at least two communication signals are both transmitted wirelessly.

26. The method of claim 18, wherein the at least two communication signals are both transmitted through a wire medium.

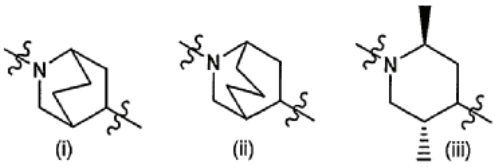
27. The method of claim 18, wherein the at least two communication signals are transmitted through a wire medium, and wirelessly.

This example does not respect paragraphs 34 and 36. As a result, the OCR engine cannot distinguish correctly word boundaries and the result of the OCR is totally unusable.

17. Bad stamp by receiving office before scanning

reagent such as diborane or alkylation of the piperidine nitrogen with an alkyl halide or sulfonate ester provides the desired compounds.

[00176] Additionally, compounds of formulae (I, Ia, and Ib) in which the piperidine ring is replaced by:



This example does not respect paragraph 41. As a result, the first six words of the text of the page cannot be read by the OCR procedure. Moreover, the stamp introduces extra invalid characters that will pollute the indexation engines if the page is not quality checked by an operator.

18. Another page with mathematical formulae badly laid out

probability of the statistical outlier event of a noise only FFT bin magnitude being larger than a FFT bin containing both signal and noise is negligible.

Define,

$$\alpha = \sum_{n=0}^{N-1} r[n]\exp[-j2\pi(\frac{\hat{f}}{f_s} - \frac{1}{2N})n] \quad (9)$$

$$\beta = \sum_{n=0}^{N-1} r[n]\exp[-j2\pi(\frac{\hat{f}}{f_s} + \frac{1}{2N})n] \quad (10)$$

Then the discriminant, or distance metric, of frequency estimation error is defined as,

$$D(\varepsilon,\hat{\varepsilon}) = \frac{|\beta| - |\alpha|}{|\beta| + |\alpha|} \quad (11)$$

$$\text{where, } \varepsilon = fT_s - \frac{k_{max}}{N} \quad (12)$$

and,

$$\hat{\varepsilon} = \hat{f}T_S - \frac{k_{max}}{N}$$

For the initial frequency estimate using the FFT, $\hat{f}_0 T_S = \frac{k_{max}}{N}$ and $\hat{\varepsilon} = 0$.

In the noiseless case,

$$D(\varepsilon,\hat{\varepsilon}) = \begin{cases} -1, & \varepsilon - \hat{\varepsilon} = \frac{-1}{2N} \\ 0, & \varepsilon - \hat{\varepsilon} = 0, \\ 1, & \varepsilon - \hat{\varepsilon} = \frac{1}{2N} \end{cases} \quad (13)$$

$D(\varepsilon,\hat{\varepsilon})$ is a monotonically increasing function of $\varepsilon - \hat{\varepsilon}$. Therefore, each $D(\varepsilon,\hat{\varepsilon})$, there is a unique inverse mapping to $\varepsilon - \hat{\varepsilon}$. Clearly, $D(\varepsilon,\hat{\varepsilon}))$ may be used as a discriminant for fine frequency interpolation between FFT bin center frequencies.

There exists some functional relationship such that,

$$\hat{f}_1 T_s = \frac{k_{max}}{N} + \psi[D(\varepsilon,\hat{\varepsilon})] \quad (14)$$

where, $\psi(.)$ is a monotone increasing function. $\psi(.)$ is called the frequency interpolation function and $\hat{f}_1$ is the first interpolated frequency estimate.

The requirement that $\hat{f}_1$ has zero error in the noiseless case is, $\psi[D(\varepsilon,\hat{\varepsilon})] = \varepsilon - \hat{\varepsilon}$, for $-1 \le D \le 1$. Therefore, $\psi^{-1}(\varepsilon - \hat{\varepsilon}) = D(\varepsilon,\hat{\varepsilon})$.

THE FREQUENCY INTERPOLATION FUNCTION

As this page does not respect many recommendations, the result of the OCR is not usable:

- Embedded mathematical formulae not separated from text paragraphs (paragraph 22)
- Unusual characters in text paragraphs (paragraph 35)
- Italic style combined with Greek characters (paragraph 38)

The recommended way to lay out this page is to use extra spaces to separate embedded formulae from the paragraphs. Greek letters should not be italicized in formulae and paragraphs. Hats (^) shall be avoided to denote variables in text paragraphs when possible; superscripts may be used instead: “epsilon hat” could be represented ε^ or ε^{hat}.

19. A page with unrecommended italic and underlined characters

2-(3-{[5-Chloro-4-(1*H*-indol-3-yl)pyrimidin-2-yl]amino}piperidin-1-yl)-*N*-methylacetamide (S Enantiomer)

LCMS 399/401 $[M+H]^+$, RT 1.88 min.

EXAMPLE 320

3-{[5-Chloro-4-(1*H*-indol-3-yl)pyrimidin-2-yl]amino}-*N*-isopropylpiperidine-1-carboxamide (Enantiomer 1)

LCMS 413/415 $[M+H]^+$, RT 3.20 min.

EXAMPLE 321

3-{[5-Chloro-4-(1*H*-indol-3-yl)pyrimidin-2-yl]amino}-*N*-isopropylpiperidine-1-carboxamide (Enantiomer 2)

LCMS 413/415 $[M+H]^+$, RT 3.19 min.

EXAMPLE 322

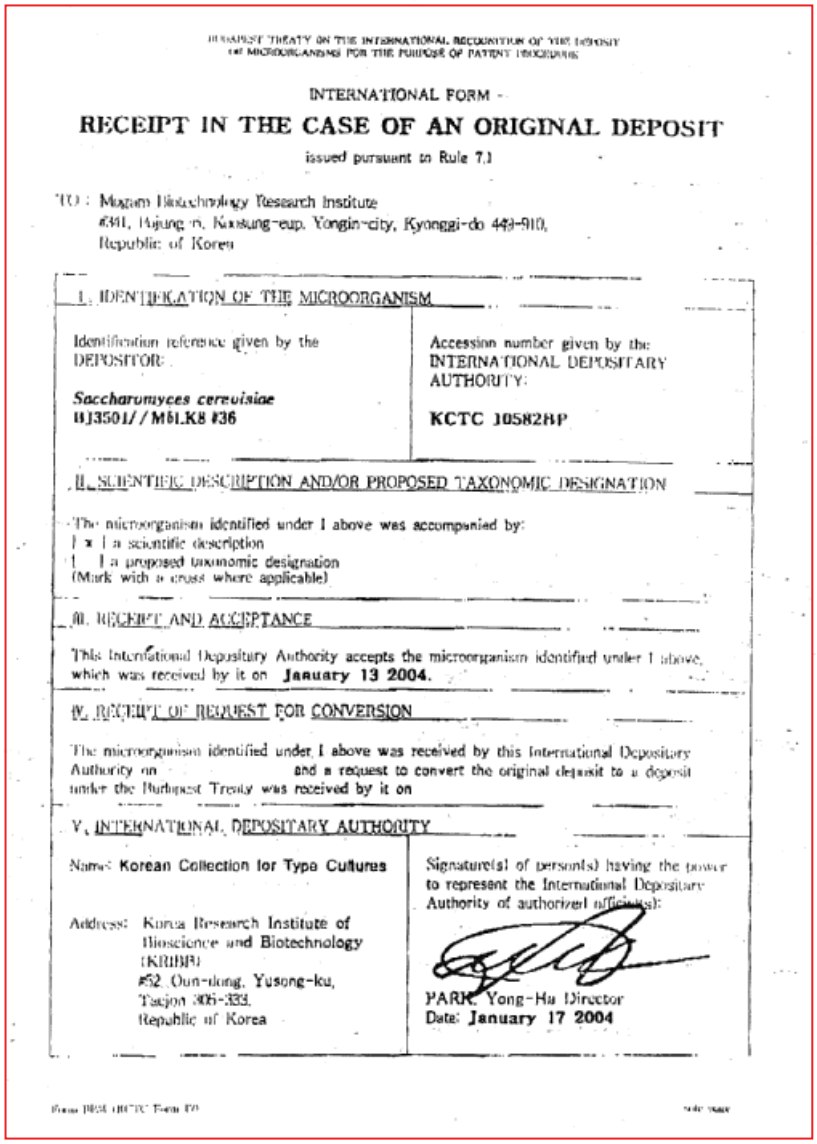
2-{3-[(4-{[5-Chloro-4-(1*H*-indol-3-yl)pyrimidin-2-yl]amino}piperidin-1-yl)carbonyl]pyrrolidin-1-yl}-*N*-methylacetamide (Racemate)

LCMS (pH 5.8) 496/498 $[M+H]^+$, RT 2.79 min.

This is a recurrent example of page with OCR problems encountered in the PCT publication. This page does not respect the following rules:

- Paragraph 37: text should not be underlined. Underlining is especially unrecommended for chemical formulae (dictionaries cannot help in these cases). Notably, this causes problems with all characters that intersect with the underline:]) y p … are not recognized correctly.
- Paragraph 38: italic style is not recommended. It is specially unrecommended to change the font style within a word (OCR engines assume often that all characters of a word share the same style). As a result, all the "1*H*" and "-*N*-" are badly recognized.

20. A page completely unreadable



BUDAPEST TREATY ON THE INTERNATIONAL RECOGNITION OF THE DEPOSIT OF MICROORGANISMS FOR THE PURPOSE OF PATENT PROCEDURE

INTERNATIONAL FORM

RECEIPT IN THE CASE OF AN ORIGINAL DEPOSIT

issued pursuant to Rule 7.1

TO: Mogam Biotechnology Research Institute
#341, Pajung-ri, Kiheung-eup, Yongin-city, Kyonggi-do 449-910,
Republic of Korea

I. IDENTIFICATION OF THE MICROORGANISM

Identification reference given by the DEPOSITOR:	Accession number given by the INTERNATIONAL DEPOSITARY AUTHORITY:
Saccharomyces cerevisiae **BJ3501//M6LK8 #36**	**KCTC 10582BP**

II. SCIENTIFIC DESCRIPTION AND/OR PROPOSED TAXONOMIC DESIGNATION

The microorganism identified under I above was accompanied by:
[x] a scientific description
[ ] a proposed taxonomic designation
(Mark with a cross where applicable)

III. RECEIPT AND ACCEPTANCE

This International Depositary Authority accepts the microorganism identified under I above, which was received by it on **January 13 2004.**

IV. RECEIPT OF REQUEST FOR CONVERSION

The microorganism identified under I above was received by this International Depositary Authority on and a request to convert the original deposit to a deposit under the Budapest Treaty was received by it on

V. INTERNATIONAL DEPOSITARY AUTHORITY

Name: **Korean Collection for Type Cultures**	Signature(s) of person(s) having the power to represent the International Depositary Authority or authorized official(s):
Address: Korea Research Institute of Bioscience and Biotechnology (KRIBB) #52, Oun-dong, Yusong-ku, Taejon 305-333, Republic of Korea	PARK, Yong-Ha Director Date: **January 17 2004**

Form BP/4 (IBTC Form IP)

This page should not be accepted by offices; it has been sent by fax at 200 dpi and is not readable. In order to deal with these cases, operators should declare the whole content of the page as an image as no text is extractable.

[End of Appendix 2 and of document]