STANDARD ST.22

DECEMBER 2008 CHANGES

On November 21, 2008, at its tenth session, the SCIT Standards and Documentation Working Group adopted a revision of WIPO Standard ST.22. As the new version was rather different and longer than the former version, instead of showing the changes made in the new version with respect to the previous version, both versions are reproduced below for information purposes.

Version of WIPO Standard ST.22 adopted on November 21

Outdated version of WIPO Standard ST.22

STANDARD ST.22

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

Revision adopted by the SCIT Standards and Documentation Working Group at its tenth session on November 21, 2008

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. Note that this document does not provide detailed recommendations for the Japanese and Korean languages; the percentage of the number of full text electronic filings to the total number of filings by year is indeed above 90% in these countries, meaning that this Standard is not applicable for the Japanese and Korean languages in practice.

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.

DEFINITIONS

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 subscripts/superscripts or if it contains the sum, integral or product mathematical symbols.

7. A bounding box of a character/set of characters is the smallest axis-aligned rectangle which includes all parts of the character/set of characters.

8. The term "cursive" refers to a stylized form of handwriting whereby the letters in words are connected, making a word one single complex stroke. Fonts are said to be cursive if they are designed to resemble handwriting.

CREATION OF THE ORIGINAL

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

PAPER SUPPORT IF FILED ON PAPER

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

    (a) The paper should be strong, white and clean.

    (b) The paper weight should be between 70, preferably 80, and 120 g/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 smearing of the ink (for example when using an ink jet printer).

PAGE LAYOUT RECOMMENDATIONS

11. Double-sided printing should be avoided. If this is not possible, sufficiently opaque paper should be used to ensure clean recto/verso scanning.

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

13. A minimum margin of 2 centimeters should be present at the top, bottom and sides of each sheet, and a minimum margin of 2.5 centimeters 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.

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

15. Page numbering should be indicated with 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.

16. 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 previously mentioned parts of the application should specify the corresponding part (in the language of the application); the claims paragraph should be numbered sequentially. The format of the claims numbering should allow for a clean separation between the claim number and the claim text for each claim. Recommended formats are either to use Arabic numerals followed by a point or to use the word "Claim" –or the equivalent in the language of the patent application-, followed by a space and the claim Arabic number, the following text of the claim being right-indented with respect to the claim number of at least 1 cm in both cases.

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

18. Pages containing paragraphs should have a portrait orientation.

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

20. Any page should contain only one direction of text.

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

22. It is recommended to avoid the use of footnotes, margin texts and headers, except as indicated in paragraph 14 (line numbering), paragraph 15 (page numbering) and for the inclusion of an applicant's file reference in the top left-hand corner of the margin.

PARAGRAPH LAYOUT RECOMMENDATIONS

23. It is recommended that tables, complex chemical formulae, complex mathematical formulae, images and drawings be separated from text paragraphs. It is advised that such items be surrounded by top and bottom blank margins of at least 1 cm that encompass the width of the page.

24. Images and drawings should at maximum be included in the "Drawings" section and referred to in the "Description" and "Claims" sections of the patent application.

25. Images and drawings should be in black and white (grayscale images should be avoided as information is lost when scanning them or converting them to black and white). Figures should contain clear lines that are thick enough to be well represented at a 300 dpi resolution.

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

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

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

29. All characters within a paragraph line should have their baselines carefully aligned, except for subscript and superscript characters as indicated in paragraph 35.

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

31. When possible, word splitting by the use of hyphens should be avoided (for example, at the end of lines or table cells). This does not apply for languages that use compound nouns (for example the German language).

TABLE RECOMMENDATIONS

32. Only white background should be used.

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

FONT RECOMMENDATIONS

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

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

36. The recommended fonts are the following in order:

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

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

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

However, the Arial, Helvetica, DejaVu Sans, Free Times 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.

37. The characters of the fonts should be well shaped, with no shadows. The spaces between characters should be wide enough (narrow spacing should be avoided).

38. Narrow and cursive fonts should not be used.

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

40. Unusual (non-standard /non-typical/ irregular) characters should be avoided as much as possible. If necessary, they should be of 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 its requirements for the character set which can be used for the preparation of the patent applications.

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

RECOMMENDATIONS FOR COMBINATION OF LANGUAGES

42. Within sections/pages of patent applications, the mixing of Asian (i.e., ideogram based) and European (i.e., Latin and Cyrillic alphabets) languages is problematic for the OCR procedures and should be avoided, except where necessary.

SCANNING RECOMMENDATIONS

43. Patent applications should be scanned either in black and white or grayscale.

44. The preferred resolution for the scanning is 300 dpi. Scanning at resolutions lower than 300 dpi, even in grayscale, can result in poor quality documents published by offices, since exchange of documents between offices and the publication processes often involve conversions to 300 dpi black and white TIFF group IV pages.

45. Scanned documents should be converted either into PDF or TIFF formats.

CORRECTIONS

46. 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. Making corrections by means of 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 standard fax resolution: pages should either be sent physically or by fax using a resolution higher than or equal to 300 dpi or by any network transfer means supported by the office, on condition that each page has been scanned at a resolution higher than or equal to 300 dpi .

RECOMMENDATIONS FOR OFFICES

47. Patent offices should avoid altering the received pages before submitting them for 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 shall take measures to ensure that the changes only occur in the margins of the documents, as defined in Appendix 1.

48. In the future, patent offices should avoid designing paper forms to handle the communication between the applicants and the office. According to past experience, designing and putting in place secured on-line forms systems is preferable to building systems to recognize paper forms. Nevertheless, the following recommendations are made for the design of paper forms in the view of facilitating their recognition:

- Lines of small dots should not be used in forms to indicate to the user where text should be entered.
- Drop-out colors should be used for the character boxes (light gray).
- Drop-out colors should not be used for areas deemed to receive grayscale or colored contents like a scanned signature or a drawing.

IMPLEMENTATION

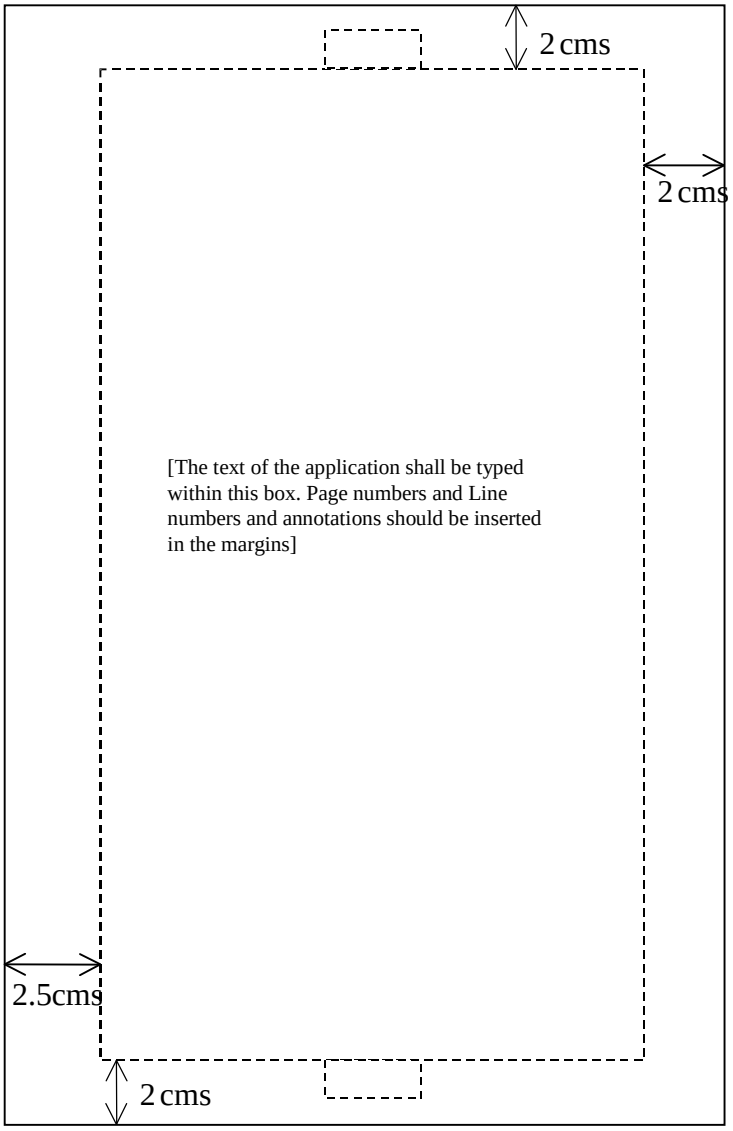
49. It is recommended that Offices intending to start accepting or requesting the filing of patent applications typed in OCR format publish full guidance in their Official Gazettes at regular intervals and in their websites, defining therein the exact character type(s) permitted, and specifying the exact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

Examples

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

[Appendices follow]

APPENDIX 1



Original Size = A4

APPENDIX 2

EXAMPLES OF GOOD AND BAD PRACTICES

You will find in this Appendix good examples and bad examples of patent document pages with respect to the accuracy obtained when performing OCR operations on them.

Examples of good practices

Example 1: a good description page

WO 2006/111319 PCT/EP2006/003401

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.

Appendix 2, page 2

Example 2: a good claims page

CLAIMS

1. A sports or games apparatus (10), characterized in that it comprises a flexible grid (11) of rectangular shape, which has two base sides (12) and two height sides (13), and is formed from a plurality of grid elements (20, 30, 40), in which the said grid elements comprise a pair of rigid bars (20) forming respectively the said two height sides of the grid, capable of being attached to corresponding support structures (S); a pair of elastic base cords or tapes (30), forming respectively the said two base sides of the grid, with their ends (31) attached to the rigid bars; and a plurality of elastic grid cords or tapes (40), the ends of each of the said elastic grid cords or tapes being attached to another two corresponding grid elements, in such a way that, in an installed condition of the said apparatus, in which the said rigid bars are attached to the said support structures in such a way as to bring the said elastic base cords or tapes and the said elastic grid cords or tapes into tension, the said grid elements are positioned so as to form a grid with a rectangular mesh.

2. An apparatus according to Claim 1, in which the ends (31, 41) of each elastic base cord or tape (30) and of each elastic grid cord or tape (40) are movable along the direction of extension of the corresponding grid element to which they are attached.

3. An apparatus according to Claim 2, in which the ends (31, 41) of each elastic base cord or tape (30) and of each elastic grid cord or tape (40) are bent back to form a noose and attached to themselves, and the corresponding grid elements to which they are attached in a movable way are inserted into the corresponding nooses (42) formed by them.

4. An apparatus according to any one of the preceding claims, in which at least one elastic grid cord or tape (40) comprises at least one intermediate point (43) attached to another elastic grid cord or tape (40).

5. An apparatus according to Claim 4, in which at least one intermediate point is movable along the direction of extension of the corresponding elastic grid cord or tape (40) to which it is attached.

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

Appendix 2, page 3

Example 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:

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

[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.

Appendix 2, page 4

Examples of bad practices

Example 1: a poor quality page with many deficiencies

WO 2005/060413 PCT/US2004/033203

LOGICAL VIEW REPORT 604115530.100802

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 Service

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 representating a Java applet. In our design the applet probably is implemented within the Silver Stream application server development and run-time environment.

Java Bean

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.

Page 9

78

This example does not conform to paragraph 10 (the page was probably submitted by fax at 200 dpi to the office – see the noise – and some text appears on heavy gray backgrounds). Nor does the example comply with paragraphs 13 and 47: a reference number (604115530.100802) is stamped on the body of the page (it should be in the margins). The page numbering is incorrect (should be "9", not "page 9", see paragraph 34). Finally, the font size is too small (paragraph 15). Such pages should ideally not be accepted by offices and replacement pages should be requested (this page is impossible to OCR correctly).

Appendix 2, page 5

Example 2: a page with a non-white background

WO 2005/097403 PCT/FR2005/050194

- 13 -

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).

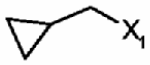
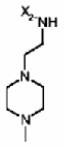
4. Dispositif (D) selon la revendication 1, CARACTÉRISÉ PAR LE FAIT QUE le bâti (210) du dispositif porte-pièce (200) est lui-même monté mobile en rotation selon un axe (B) perpendiculaire à l'axe (A) de rotation défini par les deux paliers (220 et 230) qu'il supporte.

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

This example does not conform to paragraph 12. The page needs to be filtered to attempt to remove the noisy background before submitting it for an OCR operation. If OCRed as is, the obtained text is unreadable.

Appendix 2, page 6

Example 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 300 dpi black and white in characters which are not solid. As a result, the accuracy of the OCRed text is poor (this example does not conform to paragraph 12).

Example 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 SWAYBAR 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.

Appendix 2, page 7

Example 5: a page with a non-recommended layout and other deficiencies

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.

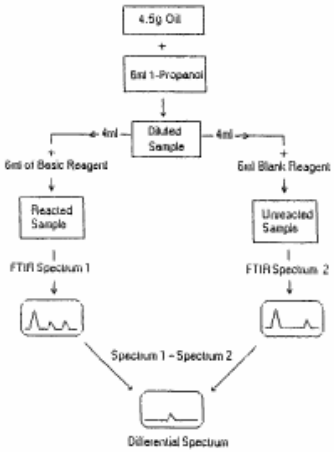


Figure 2. Analytical protocol for the determination of AN by FTIR spectroscopy.

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.

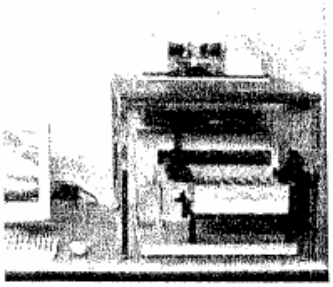


Figure 3. The COAT AN/BN Analyzer and its key components

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 conform to the following recommendations: paragraph 17 (single column formatting), paragraph 39 (uses italic and bold fonts), paragraph 46 (has manual corrections performed after printing). The left-right justification of the paragraph is also not recommended (paragraph 30), although in this case, this would not have negative effects on the OCR since the words are still sufficiently separated by white spaces. Nor, finally, does the example comply with paragraph 27 (one and a half line spacing).

Appendix 2, page 8

Example 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 for the OCR engines for several reasons (paragraph 14):

- they may not be aligned with the lines they correspond to, leading to baseline detection defaults;
- they could be too small, resulting in 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 35).

Appendix 2, page 9

Example 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 conform to paragraph 20.

One of the limitations of the best OCR engines available today is that they can read only one direction of text on one page (a preprocess of the page is to detect the main text orientation of the page). As a result, all of 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, etc.).

Appendix 2, page 10

Example 8: a page with mixed embedded mathematical formulae and text

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

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

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,

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

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,

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

$$\Delta E_{mag}^{spin} = g\mu_B B \tag{37}$$

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 $n = 1$, $\ell = 0$

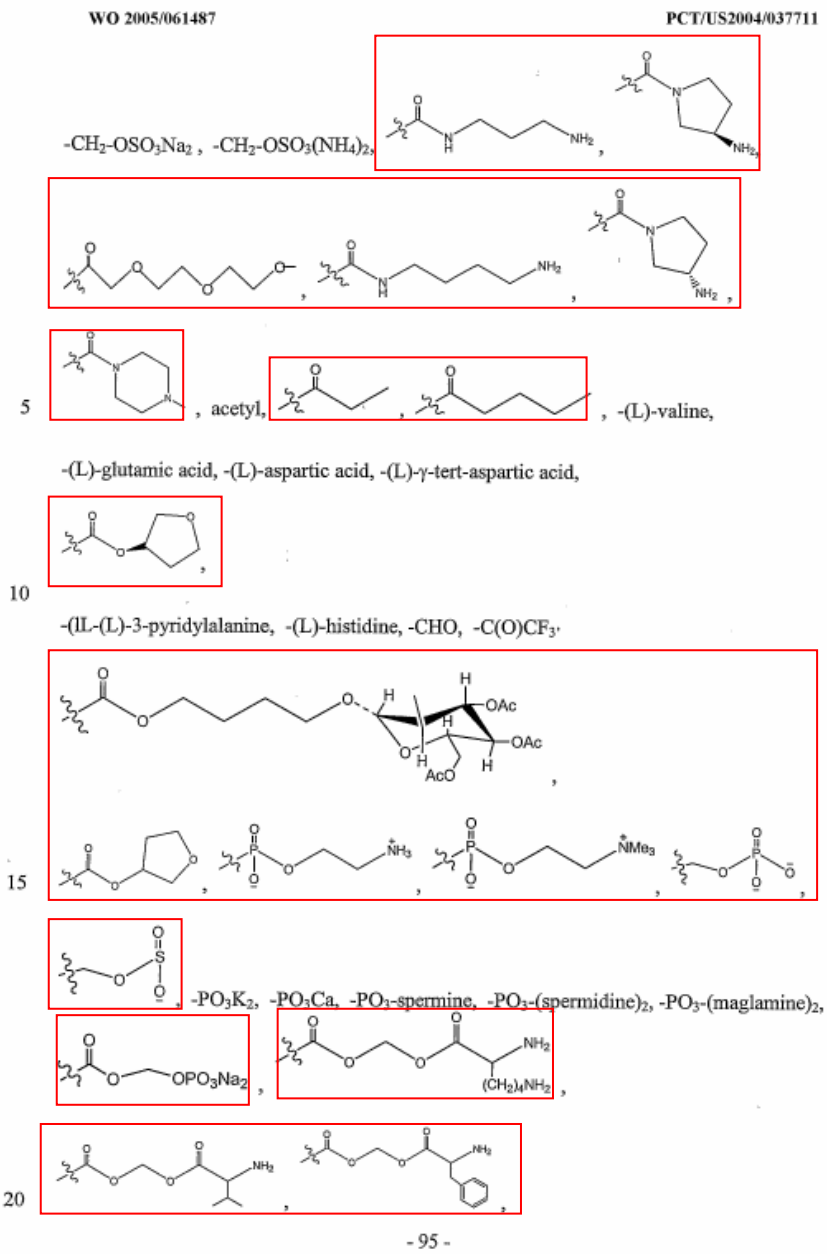
This example does not conform to paragraph 23. The OCR engine is not able to separate correctly the text and the formulae (see the result of a manual segmentation of the formulae in red: the embedded formulae even intersect).

As a general comment, in this example, the text and the formulae are too dense for good recognition; neither does the example comply with paragraphs 27 and 28.

This example also uses unusual characters: Greek symbols can be used even if they increase the recognition difficulty of the page (see paragraph 40). However, it is highly recommended not to combine italics, bold or underlined fonts with unusual characters (paragraph 39).

Appendix 2, page 11

Example 9: a page with mixed embedded chemical formulae and text



WO 2005/061487 PCT/US2004/037711

-CH_2-OSO_3Na_2 , -CH_2-$OSO_3(NH_4)_2$,

acetyl, -(L)-valine,

-(L)-glutamic acid, -(L)-aspartic acid, -(L)-γ-tert-aspartic acid,

-(IL-(L)-3-pyridylalanine, -(L)-histidine, -CHO, -C(O)CF_3,

-PO_3K_2, -PO_3Ca, -PO_3-spermine, -PO_3-(spermidine)$_2$, -PO_3-(maglamine)$_2$,

- 95 -

This example does not conform to paragraph 23. 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.

Appendix 2, page 12

Example 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)$_n$phenyl, -O(CH_2)$_n$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)$_n$(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_{3-6}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 subscript characters are too small to allow for accurate recognition. This phenomenon is frequently encountered for patents in the chemistry field.

Appendix 2, page 13

Example 11: a page with badly formatted tables

WO 2005/063765 PCT/US2004/043492

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
					$CHCH_3CH_3$
					$COOCH_2CH_3$
					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 (does not conform to paragraph 33). 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 34 and 35)
- The baselines of the column headings are mixed (paragraph 29). As a result, the engine will detect wrongly 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…

Appendix 2, page 14

Example 12: a justified page

WO 2005/087962 PCT/EP2005/002268

1

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

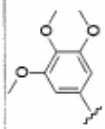
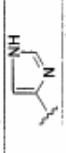
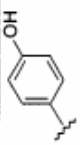
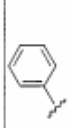
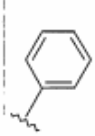
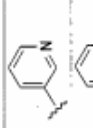
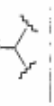
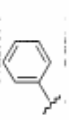
In this example, left and right justifications are applied to the paragraphs. If this makes the text more aesthetic looking, it sometimes makes OCR operations difficult when the separations between the words become too small (paragraph 30). Neither does this example conform to paragraph 31, which states 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).

Appendix 2, page 15

Example 13: a table with bad boundaries

WO 2004/110415 PCT/EP2004/051048

- 60 -

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 original received 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 undesired "junk" characters that will make the indexation of the document by search engines less effective.

Appendix 2, page 16

Example 14: bad subscript and superscript characters

WO 2005/100305 PCT/IB2005/000872

-9-

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-C_6)alkyl, (C_1-C_6)alkoxy, (C_1-C_2)alkyl substituted with one or more halogens, (C_1-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-C_6)alkyl, (C_1-C_6)alkoxy, (C_1-C_2)alkyl substituted with one or more halogens, (C_1-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

The following problems exist in this example (paragraph 35):

- 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 34 and 35 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"

Appendix 2, page 17

Example 15: an example with unusual characters

WO 2006/057705 **PCT/I**

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

The following problems exist in this example:

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

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

Appendix 2, page 18

Example 16: an example with narrow fonts and narrow spacing

WO 2006/036330 **PCT/US2005/028798**

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.

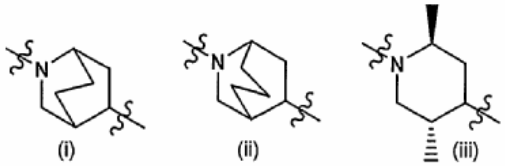
This example does not conform to paragraphs 37 and 38. As a result, the OCR engine cannot correctly distinguish word boundaries, and the result is that the OCR is totally unusable.

Appendix 2, page 19

Example 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 conform to paragraph 47. 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.

Example 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,

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

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 conform to many recommendations, the result of the OCR is not usable:

- embedded mathematical formulae not separated from text paragraphs (paragraph 23);
- unusual characters in text paragraphs (paragraph 40);
- italic style combined with Greek characters (paragraph 39).

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. Circumflexes (^) shall be avoided to denote variables in text paragraphs when possible: superscripts may be used instead: "epsilon circumflex" could be represented ε^ or $\varepsilon^{circumflex}$.

Appendix 2, page 21

Example 19: a page with italic and underlined characters

WO 2006/038001 PCT/GB2005/003827

- 132 -

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 frequent OCR problem encountered in the PCT publication. This page does not conform to the following recommendations:

- Paragraph 41: text should not be underlined. Underlining is especially not recommended for chemical formulae (dictionaries cannot help in these cases). Notably, this causes problems with all characters that intersect with the underline: `l ) y p` … are not recognized correctly.
- Paragraph 39: italic style is not recommended. It is highly recommended not to change the font style within a word (OCR engines assume often that all characters of a word have the same style). As a result, all the "1*H*" and "-*N*-" are badly recognized.

Appendix 2, page 22

Example 20: a page completely unreadable

WO 2005/071074 PCT/KR2005/000214

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: Magam Biotechnology Research Institute
#341, Pojung-ri, Kiusung-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// M61.K8 #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** Address: Korea Research Institute of Bioscience and Biotechnology (KRIBB) #52, Oun-dong, Yusong-ku, Taejon 305-333, Republic of Korea	Signature(s) of person(s) having the power to represent the International Depositary Authority of authorized official(s): PARK Yong-Ha Director Date: **January 17 2004**

Form BP/4 (IBTC Form 19)

This page should not be accepted by offices: it has been sent by fax at 100 dpi and is not even readable by the human eye. In order to deal with these cases, operators declare the whole content of the page as an image as no text is extractable.

[End of Appendix 2 and of Standard]

NOTICE: This version of WIPO Standard ST.22 is outdated and was replaced with the version adopted by the SCIT Standards and Documentation Working Group at its tenth session on November 21, 2008

OUTDATED VERSION OF STANDARD ST.22

RECOMMENDATION FOR THE PRESENTATION OF PATENT APPLICATIONS TYPED IN OPTICAL CHARACTER RECOGNITION (OCR) FORMAT

INTRODUCTION

1. This Recommendation has been established so as to assist 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.

2. 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 object 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.

3. 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 data base 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

4. 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.

CREATION OF THE TYPED ORIGINAL

5. A patent application will often be prepared using word processing equipment or various types of electronic or electric typewriters in which the type font and size can be readily selected by the interchange of a daisy-wheel or golfball print head. Experience has shown that the most efficient format of type to permit OCR equipment to be reliably used is that defined in the International Standard Organization (ISO) Standard 1073/II, the so-called OCR-B format.

Paper Support

6. 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 80 and 120 gms/m^2;

(c) the paper size should preferably be A4, viz. 210 mm x 297 mm.

Type Characteristics

7. The typing of the text should be as follows:

(a) the characters should be evenly typed in black with a sharp rendition;

(b) ribbons should be used once only;

(c) a constant character pitch of either 10 or 12 characters per inch should be used;

(d) typing should be done at one and a half line spacing;

(e) a minimum margin of *2* cms should be present at the top, bottom and sides of each sheet, and one of 2.*5* cms on the left side of each sheet. Any applicant's or representative's references should appear in the margin at the top. It is recommended that the typing paper used has a feint pre-printed guide box within which text is typed, a sample of which is given in *Appendix I* to this Annex. Line numbering, if given, should be typed using arabic characters in the left hand margin area, at least 0.*5* cm outside of the box as shown in *Appendix I*;

(f) text should not be underlined;

(g) page numbering should be given by simply using arabic characters without other delimiting character. Page numbers should preferably be centered at the top of the sheet, as shown in *Appendix I*;

(h) word splitting at the end of a line by the use of hyphens should be avoided. The right hand margin of the typed copy should not be justified;

(j) the description, the claims(s) 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 afore-mentioned parts of the application should specify the corresponding part (in the language of the application);

(k) tables, chemical and mathematical formulae should be typed in the body of the text as far as is practical. If the complexity of the tables or formulae so dictate, then they should be presented on separate sheets with suitable references thereto inserted in the text;

(l) Greek, mathematical and other characters not provided on a normal typewriter should be inserted by hand into the text; alternatively, their substitutes as recommended in the relevant international or national standards may be used. Examples of possible substitutes are given in *Appendix II* to this Annex;

(m) the use of footnotes should be avoided.

(n) the numeral key "1" should be used for the number "one", and the letter key "1" for the letter "ell". The letter key "O" should not be used for the zero sign and vice versa.

Corrections

8. Corrections to the text of an application should preferably be done by reprinting whole pages. For the purposes of making corrections, the use of white correcting fluid or self adhesive strips of paper should be avoided.

FILING OF TYPED ORIGINAL

9. The typed original should be filed in a strong envelope, preferably of transparent plastic. The typed sheets should be free of creases and should not be rolled.

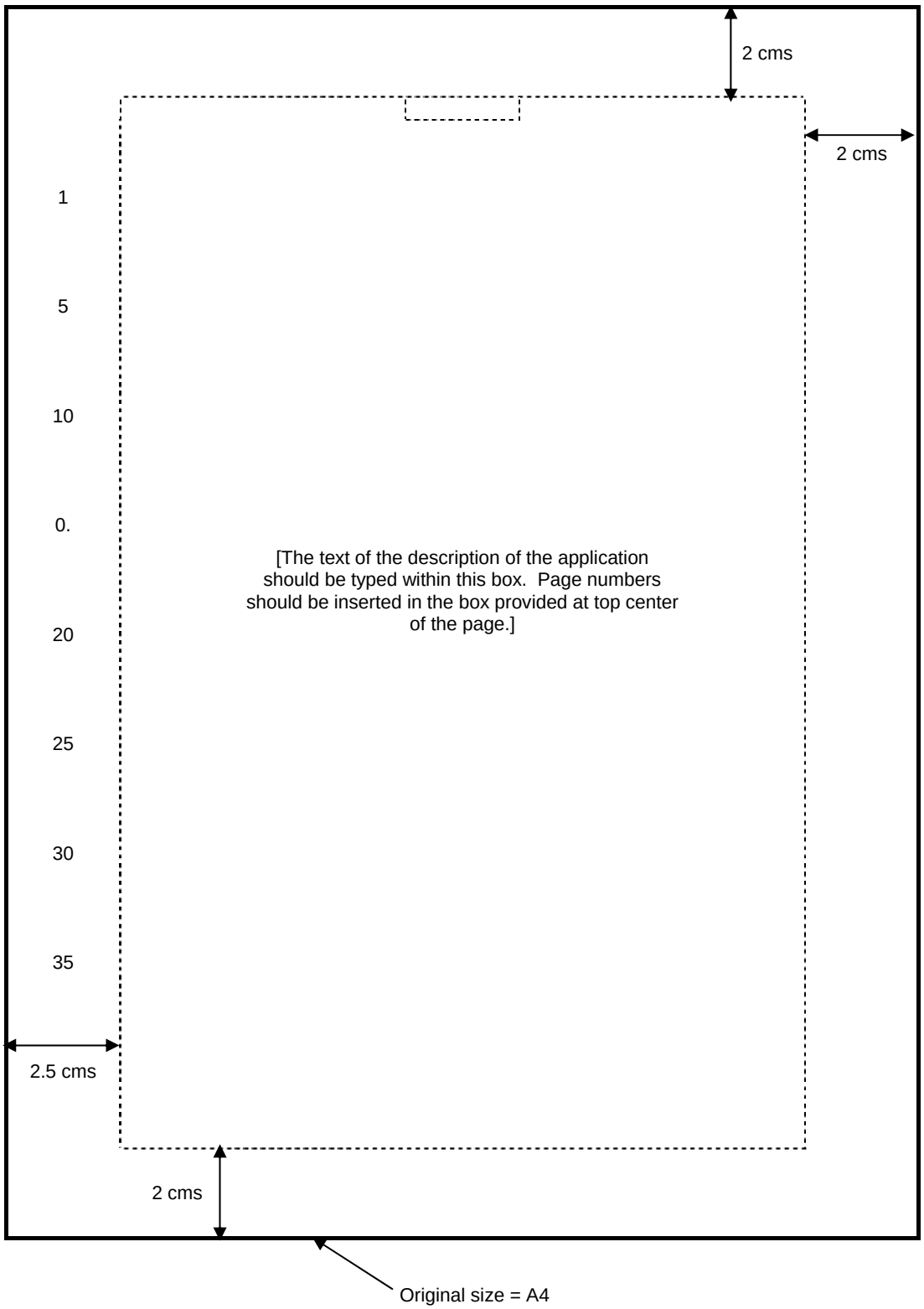
IMPLEMENTATION

10. 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, defining therein the exact character type or types permitted, and specifying the exact paper size allowable.

[Appendices follow]

NOTICE: This version of WIPO Standard ST.22 is outdated and was replaced with the version adopted by the SCIT Standards and Documentation Working Group at its tenth session on November 21, 2008

APPENDIX I



[Appendix II follows]

NOTICE: This version of WIPO Standard ST.22 is outdated and was replaced with the version adopted by the SCIT Standards and Documentation Working Group at its tenth session on November 21, 2008

APPENDIX II

EXAMPLES OF POSSIBLE SUBSTITUTES
FOR SPECIAL CHARACTERS

Special character		Substitute
Α	α	Alpha
Β	β	Beta
Γ	γ	Gamma
Δ	δ	Delta
Ε	ε	Epsilon
Ζ	ζ	Zeta
Η	η	Eta
Θ	ϕ	Theta
Ι	ι	Iota
Κ	κ	Kappa
Λ	λ	Lambda
Μ	μ	My
Ν	ν	Ny
Ξ	ξ	Xi
Ο	ο	Omicron
Π	π	pi
Ρ	ρ	Rho
Σ	σ	Sigma
	Ζ	Sigma
Τ	τ	Tau
Υ	υ	Ypsilon
Φ	φ	Phi
Χ	χ	Chi
Ψ	ψ	Psi
Ω	ω	Omega
	½	1/2
œ		oe
‰		Promille/parts per thousand
©		c.

[End of Appendix II and of Standard]