

WIPO



WO/PBC/4/2 Add.1

ORIGINAL: English

DATE: September 10, 2001

WORLD INTELLECTUAL PROPERTY ORGANIZATION

GENEVA

E

PROGRAM AND BUDGET COMMITTEE

Fourth Session

Geneva, September 17 to 19, 2001

CONTINUATION OF THE PCT OCR PROJECT

Document prepared by the Secretariat

INTRODUCTION

1. The present document is prepared by the Secretariat in response to a recommendation made by the Information Technology Projects Working Group (ITPWG) of the Standing Committee on Information Technologies (SCIT) at its first session held from September 3 to 7, 2001.
2. At its third session, held from April 25 to 27, 2001, the Program and Budget Committee (PBC) noted the contents of document WO/PBC/3/3 on "Information Technology projects to be financed by surplus resources," and requested that the text be submitted to the first meeting of ITPWG for "technical review." The document contained proposals for five major IT projects to be implemented in the 2002-2003 biennium with a total budget of 99,547,000 Swiss francs. This amount consisted of 82,900,000 million Swiss francs to be funded by a new appropriation from the Special Reserve Fund and the roll-over into the 2002-2003 biennium of 16,647,000 Swiss francs representing the unspent balance of the IMPACT (PCT automation) Project budget.
3. In discussing document WO/PBC/3/3, the ITPWG agreed that the five projects listed (WIPONET, IMPACT, PCT Electronic-filing, CLAIMS and AIMS) were all of the highest priority and that the budgets allocated to each activity were justified (see document SCT/ITPWG/1/13).
4. However, the Working Group was concerned that the Information Technology Program for the 2002-2003 biennium did not include provision for the continuation of the project for the Optical Character Recognition (OCR) of published PCT applications. This activity was

considered of sufficient importance that the ITPWG requested the International Bureau to propose to the PBC an increase in the budget of the Information Technology Program (Main Program 15), of an amount of 4 million Swiss francs, to fund this work in the next biennium. Such a proposal was to include detailed estimates for the cost of the work as well as a justification for the activity based on business benefits to users of the patent system and the technical information it generates.

BUSINESS JUSTIFICATION

Use by Patent Offices

5. The most important task of patent offices (POs) is the granting of valid patents. Of paramount importance in the patenting process is the ability to conduct a thorough search of the prior art. To carry out prior art searches, POs are responsible for creating and maintaining search files containing over 30 million patent documents. With the rapid growth in the size and complexity of this documentation, accessing patent documents in a timely, cost efficient manner is necessary to ensure high quality search results.

6. Currently there are two ways that most POs gain access to the documents in the prior art. The first is to place the patent documents in search files using a patent classification system, to allow easy access for searchers. The most commonly used of these classification systems is the International Patent Classification (IPC), but the IPC by itself cannot provide all the access necessary to ensure the highest quality search. Therefore access to patent documents is also carried out needed by the use of full text searching. Most patent offices, as part of their patent publication process, create a file of their published patent documents in character readable and searchable format. This is not the case for the published PCT international applications.

7. Under the current PCT OCR Project, these published PCT applications are converted (for the most part) into character format immediately after publication and made available in WIPO's Intellectual Property Digital Library (IPDL) system as well as to other Patent Offices for inclusion in their own internal databases. Clearly, with more than 200,000 PCT international applications expected to be published in the next biennium, the loss of this character coded data for facilitating access to the PCT published applications will have a very negative effect on the quality of the search process.

Use by the General Public and Other Interested Parties

8. The information contained in PCT applications is of particular importance to many individuals and companies who are involved in research and development. This data has enormous value in that it can shorten the research and development cycle by providing valuable information immediately into the hands of the users of such information free of charge. This position is supported by the heavy use made of WIPO's IPDL web site system, which receives millions of enquiries per week.

9. If the PCT OCR scanning activity were to be discontinued, then the global users of this information, which are currently relying on it in their research and development work, would be deprived of a very important resource that not only supports innovation generally but also encourages innovation at the national level, especially in Developing Countries.

10. It should also be emphasized the IPDL system adds considerable value to the WIPONET content, which among other things has as one of its goals, the distribution of information to national Intellectual Property Offices.

REPORT OF THE ITPWG DISCUSSION

11. The issue of the continuation of the PCT OCR Project was initially raised at the ITPWG in the context of the cessation, at the end of 2001, of a stand-alone Intellectual Property Digital Libraries (IPDL) project. After this date the IPDL system will go into production for WIPO data collections and work on technical standards for digital libraries will be subsumed into the work plan of the Information Technology Standards and Documentation Section. Currently this project is responsible for the outsourced PCT OCR activity and to-date has funded it from the project budget.

12. Many delegations took the floor to stress the importance of the electronic data capture of PCT applications, not just for intellectual property Offices (IPOs) but also to individual users of the patent system and the patenting community as a whole, as it provided much needed technical information in a user-friendly and searchable full-text format. The discussion on this matter is covered in the report of the ITPWG meeting, document SCIT/ITPWG/1/13, paragraphs 124 to 128.

BUDGET IMPLICATIONS

13. For the last three years the WIPO IT program has funded the OCR conversion of PCT applications into full text using an external company. However due to the priorities established for the IT program for the 2002-2003 biennium, the budget of the Information Technology Program, as it is currently proposed, does not have sufficient funds to continue this activity.

14. During the ITPWG, the Secretariat was able to give provisional estimates for the continuation of the PCT OCR activity in the 2002-2003 biennium. These consist of fixed cost elements for quality control (150,000 Swiss francs) and development costs for the loading of data and format checking when the data arrives at WIPO (250,000 Swiss francs). Production costs of 1.25 and 1.55 million Swiss francs were also estimated for the years 2002 and 2003 respectively. However, these costs are variable and are calculated on a per page basis, therefore, they are likely to rise with the increase in applications to the PCT. Also, as this activity is largely outsourced, funds will be required to process the renewal of the contract for the work. With this in mind the ITPWG recommended that a budget limit of 4 million Swiss francs be set for the activity for the whole of the next biennium.

15. As has already been noted, the work of the IPDL project, which currently funds this activity, will be subsumed into the Information Technology Services Division from January 2002. Therefore, the Director General has proposed the creation of a new sub-program 15.8, for PCT OCR, to be added to the Main Program 15 for Information Technology. Annex I to this document contains the text for the new sub-program and Annex II a revised Table C for Main Program 15. This 4 million Swiss francs project would be financed by the PCT Union and would increase the object of expenditure 'Publishing.' The proposed 2002-2003 budget for Main Program 15 increases from 148,656,000 Swiss francs to 152,656,000 Swiss francs and WIPO's budget increases from 678,400,000 Swiss francs to 682,400,000 Swiss francs.

15. The Program and Budget Committee is invited to give its comments on and consider recommending to the Assemblies:

(a) the inclusion of the PCT OCR activity as a sub-program (15.8) within Main Program 15, Information Technology, of the draft Program and Budget for 2002-2003;

(b) a corresponding budgetary allocation to sub-program 15.8, PCT OCR, of an amount of 4 million Swiss francs, increasing WIPO's proposed budget for 2002-2003 to 682,400,000 Swiss francs;

(c) that this amount for PCT OCR be financed by the PCT Union, thus reducing the RWCF level by the same amount.

[Annexes follow]

ANNEX I

Draft Text for Sub-program 15.8

SUB-PROGRAM 15.8
PCT OCR

Objective:

- ◆ To capture the full-text of PCT applications using Optical Character Recognition (OCR) techniques and to make that data available to Member State Offices and the WIPO Intellectual Property Digital Libraries in electronic form

Background

Having access to the technical information contained in PCT applications is critical to Patent Offices in their search and examination activities.

The most important task of patent offices (POs) is the granting of valid patents. Of paramount importance in the patenting process is the ability to conduct a thorough search of the prior art. To carry out prior art searches, POs are responsible for creating and maintaining search files containing over 30 million patent documents. With the rapid growth in the size and complexity of this documentation, accessing patent documents in a timely, cost efficient manner is necessary to ensure high quality search results.

Currently there are two ways that most POs gain access to the documents in the prior art. The first is to place the patent documents in search files using a patent classification system, to allow easy access for searchers. The most commonly used of these classification systems is the International Patent Classification (IPC), but the IPC by itself cannot provide all the access necessary to ensure the highest quality search. Therefore access to patent documents is also carried out needed by the use of full text searching. Most patent offices, as part of their patent publication process, create a file of their published patent documents in character readable and searchable format. This is not the case for the published PCT international applications.

Under the existing PCT OCR Project, these published PCT applications are converted (for the most part) into character format immediately after publication and made available in WIPO's Intellectual Property Digital Library (IPDL) system as well as to other Patent Offices for inclusion in their own internal databases. Clearly, with more than 200,000 PCT international applications expected to be published in the next biennium, the loss of this character coded data for facilitating access to the PCT published applications will have a very negative effect on the quality of the search process.

A second important use of this information is that which is made by the many individuals and companies involved in research and development activities. This data has enormous value in that it can shorten the research and development cycle by providing immediate free access to the technical information required in this process. This position is supported by the heavy use made of WIPO's IPDL web site system, which receives millions of inquiries per week. Clearly, this information is a very important resource that not only supports innovation

generally but also encourages innovation at the national level, especially in Developing Countries.

It should also be emphasized the IPDL system adds considerable value to the WIPONET content, which among other things has as one of its goals, the distribution of information to national Intellectual Property Offices and the Intellectual Property community at large.

Expected Results	Performance Indicators
1. Availability of the full-text of PCT applications in a timely manner	ø The electronic data should be delivered and uploaded according to schedule
2. The electronic data will be of a sufficiently high quality for use by Patent Offices for search and examination	ø The quality of the electronic data should meet WIPO's specifications
3. The IPDL system will continue to be heavily used	ø Number of inquiries made on the IPDL web site

Activities

- ◆ Establish and manage a contract with an external supplier for the provision of the full-text of PCT applications.
- ◆ Ensure that the electronic data provided meets WIPO's quality specifications.
- ◆ Load the electronic data into the WIPO IPDL system in a timely manner.

[Annex II follows]

ANNEX II

Revised Table C showing Main Program 15 Budget Allocation by Sub-program and Detailed Object of Expenditure, including the PCT OCR Project

C. Budget Allocation by Sub-program and Detailed Object of Expenditure
(in thousands of Swiss francs)

<i>Object of expenditure</i>	<i>Sub-program</i>								<i>Total</i>
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	
Staff Expenses									
Posts	1,432	17,957	2,742	5,628	1,760	734	362	-	30,615
Short-term Expenses	150	384	-	-	-	-	-	-	534
Travel and Fellowships									
Staff Missions	190	400	500	500	413	60	500	-	2,563
Government Officials	550	-	130	-	-	-	-	-	680
Contractual Services									
Conferences	360	120	-	50	473	20	-	-	1,023
Consultants	40	2,300	1,139	2,660	-	1,734	6,000	-	13,873
Publishing	-	50	20	-	-	-	-	4,000	4,070
Other	60	13,519	23,607	12,909	11,472	-	1,000	-	62,567
Operating Expenses									
Premises and Maintenance	-	2,600	622	15,000	616	112	300	-	19,250
Communication and Other	40	580	315	100	-	-	200	-	1,235
Equipment and Supplies									
Furniture and Equipment	-	5,200	75	1,000	3,566	140	1,438	-	11,419
Supplies and Materials	-	3,177	150	1,000	-	400	100	-	4,827
Total	2,822	46,287	29,300	38,847	18,300	3,200	9,900	4,000	152,656

[End of Annexes and of document]