World Library and Information Congress: 71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme: http://www.ifla.org/IV/ifla71/Programme.htm

July 7, 2005

Code Number: **156-E**
Meeting: **97 Newspapers**

中国报纸数字化趋势

Trend of Newspaper Digitization in China

YANG BIN
Datum Data Company, Being, China

Abstract

As a representative of a leading data capture service provider in China, Datum Data Company, I am honored to be here and would like to thank IFLA for giving me the opportunity to share my thoughts with you today.

Most of our recent work has been digitization of newspapers, including many historical newspapers, and books. I will first provide an overview of the latest trend of Newspaper digitization in China. Then I will talk about technical challenges in newspaper digitalization as well as the technical expertise of Datum Data Company

去年会议上，介绍了中国报纸数字化的一些基本情况。今年主要介绍我们感觉到的发展趋势。
Continuing the theme of our presentation in this conference last year, I would like to provide an overview of the latest trend of newspaper digitization in China.

趋势之一：越来越多报纸图书馆从事数字化工作
Trend No. 1: Rapid market growth in newspaper and library digitization.

数字化正在发展中，由于已经数字化的报纸优势逐步表现出来，越来越多的报纸寻求数字化。我们一年中加工了超过50万版，30亿字的数字化项目。
With the increasing awareness of the benefits from newspaper digitization, a growing number of news organizations are willing to make the investment. During the course of last year, we digitized more than 500,000 pages with over 3

billion characters.
We believe the digitization market will continue to grow for many years to come.

趋势之二：电子数据的转换
Trend No. 2: Emerging need for digital data transformation

由于在印刷报纸的排版过程中，中国大多数的报纸、书籍都用了方正(Founder)系统，该系统是专门为了印刷而设计的，所以，在加入专门为检索而建立的数据库的过程中，有很多的工作要做。广告经常是由客户提供的，电子数据需要重新制作。
The Founder publishing system -- used by majority of Chinese newspapers and books -- is mainly designed for printing purpose, as opposed to data retrieval. Also, some newspaper advertisements provided by clients are in various formats and need to be reprocessed. Therefore, there are many technical challenges for adding the retrieval database.

趋势之三：粗精度文本的使用
Trend No. 3: Increasing usage of rough precision digital data

越来越多的客户选择PDF格式。很多采用了双层PDF(阅读图像，检索定位用文字)。在图书的制作上，DJVU格式被采用。这些说明了，低精度的文本，在发挥着重要的作用
A growing number of clients choose to use PDF format, especially Double-layer PDF (images for reading, text for retrieving); DJVU format is often adopted in book digitization. These are examples of the increasing usage of rough precision text.

趋势之四：标引信息越来越丰富
Trend No. 4: Increasing richness of annotation information

更多的客户选择丰富的标引信息。从最初的三四项，增加到10多项。下面是一个典型的标引项的列表:
More and more clients choose to use rich annotation information -- Increased from 3 - 4 annotation items previously to more than 10 annotation items. The following is a typical list of annotation items:

报名	Newspaper name	正文	Context
期号	Issue	体裁	Types
来源	Source	分类	Classification
标题	Title	插图	Illustration
副题	Subtitle	插图作者	Illustration author
作者	Author	插图说明	The illustration explaining
日期	Date	其它	Other comment
版次	Edition number		
版名	Edition of name		
栏目	Column		

趋势之五：统计被较多的采用
Trend No. 5: Increasing usage of statistical analysis.

流行语，词语被使用的频率统计受到关注
Special attention is paid to the usage frequency of popular words or phrases.

趋势之六： Internet与数据加工
Trend No. 6: Internet- based data process

由于网络越来速度越快，更多的业务从网上传输、交互。点通建立了专门的网站业务，客户可以在网上进行数据交流。
Because the high-speed Internet access is readily available, increasing number of data processing operations are conducted remotely, and data are transferred / exchanged via the Internet. Datum has established a dedicated web server for secured clients' data exchange.

点通公司已加工过的报纸(部分)
We have digitized over 80 Chinese newspapers. Here are a few examples:

香港《大公报》100年图像数据光盘
《西藏日报》45年全文数据光盘(藏、汉两种文字)
《文汇报》61年全文数据光盘
《新民晚报》70年全文数据光盘
《人民日报海外版》15年全文数据光盘
《哈尔滨日报》55年全文数据光盘
….

- *Hong Kong Ta Kung Pao* (http://www.takungpao.com)- full-text and full-image database of articles dating back to 100 years ago
- *Tibet's Daily* (http://www.tibetinfor.com)- multi-language (Chinese and Tibetan) full-text database in of articles dating back to 45 years ago

- *Shanghai Wen Hui Bao* (http://www.whb.com.cn)- full-text database of articles dating back to 61 years ago
- *Shanghai Xi Min Wan Bao* (http://www.xmwb.com.cn)- full-text database of articles dating back to 70 years ago
- *People's Daily Oversea Edition* (http://www.rhwx.com)- full-text database of articles dating back to 15 years ago
- Harbin daily (http://www.harbindaily.com)- full-text database of articles dating back to 55 years ago

…

报纸数字化的困难
Here are the top challenges in Chinese newspaper digitization

1. 版面变化复杂，字体变化复杂。
Challenge 1: complex layouts and mixed traditional and simplified Chinese characters

一份省市级的报纸，从创刊到现在，由于历史原因经历了下面几个阶段的变化：

Typical layouts of a historical regional newspaper are as follows:

50年代初——繁体竖版，版式复杂；

50年代中——繁体横排版；

56年至60年代末——繁简混合，横版为主，版式简单；

70年至80年代——铅字印刷，其中1978年有部分二次简化字，版式简单；

80年代开始——激光照排，版式变化多；

90年代末开始——各报社都保存有电子数据。

- Early 1950s — vertical stroke in traditional Chinese characters; complex format
- Mid 1950s — horizontal stroke in traditional Chinese characters
- Mid 1950s to End of 1960s —primarily horizontal stroke in mixed traditional and
- simplified Chinese characters; simple format
- 1970s to 1980s —type printing with different versions of simplified Chinese characters; simple format
- 1980s to 1990s —laser illumination; various formats
- Late 1990s to now —digitized data

2．报纸纸质变质严重，加工困难

Challenge 2: Severely damaged paper copies of historical newspapers

就浙江报业集团的主报《浙江日报》来说，创刊于1949年，经过五十多年，早期的报纸已经变质，很难翻阅，纸质变黄、破旧。在数字化加工的过程中要保护这些报纸非常困难。

Paper copies of historical newspapers are often fragile, stained, and blurry.

It is difficult to process and protect these paper copies/

3．数据量大
Challenge 3: Huge amount of data

一般的，一个省级报纸大约10万版，中间过程数据非常庞大，成品数据(包括原版图像)一般在200G左右。如果需要制作PDF，空间还要增加。

On average, we need to digitize over 100,000 pages for a regional newspaper. The size of delivered data with images is around 200GB, not to mention the huge amount of data generated during the process of digitization. If the delivered data is in PDF format, the size is even larger.

4．质量要求严格
Challenge 4: Requirement of high data accuracy from clients

随着报纸数字化的发展，对数字化的质量要求越来越严格，正文要求万分之一的差错率。除此以外，对那些重要的文章要求十万分之一甚至百万分之一的错误率，对重要的人名、地名不能出错，对标题、作者部分一般要求十万分之一的错误率。

With the development of digitization, clients′ requirements for accuracy are increasing. Error rate must be less than one ten thousandth for general context; one hundred thousandth to one millionth for important articles; one hundred thousandth for titles and names of the authors; and 0 for important name and address.

5．原始数据类性变化多样
Challenge 5: Various types and incomplete original data

对原始数据类型来说，部分数据是纸质(非电子)，部分数据是电子数据，并且格式多样，如：大样PS文件、PDF文件、txt电子数据。电子数据PDF版面内容缺漏，不完整，特别是特殊字符、图、表、广告区内容基本没有；txt电子文本标识字段混乱、文章缺漏现象严重等等问题。

Original data are usually dispersed in various media (such as paper copies or digital data) and in various formats (such as PS, PDF, txt..). Some information is often missing in the original PDF pages, such as special symbols, graphs, tables, or advertisements. Txt files are often corrupted and missing important articles.

点通的技术特色
I would like to talk about Datum′s technical expertise, which enabled us to overcome many challenges in newspaper digitization and provide excellent services to our clients.

一、OCR技术
1. OCR technology

点通是世界上少有的甚至是唯一的全面使用OCR技术进行数字化加工的服务商。自有的技术，一个精细加工的流水线数字工厂。可以集成所有的商用OCR
Datum is one of the few data capture service providers that use OCR technology in all phases of the data-processing operations. We deployed a precise pipeline data-processing factory based Datum owned OCR technology. Our operation line can also integrate with other commercial OCR software.

二、多文种加工
2. Multi-Language Digitization Process

点通可以加工各种文字，并且达到很高的精度。点通公司已经积累了非常成熟的经验，可以加工中文、英文、法文、德文等等各种文字的报纸、图书、文件。
Datum is capable of processing data in many languages with high accuracy. We have extensive experience on digitization of documentations, newspapers, and books in various languages, including Chinese, English, French, Germany, etc.

三、独特的质量控制办法通过了ISO9001:2000版质量体系认证
3. ISO9001 2000 certified quality-control process

点通有自己独创的质量控制的办法。保证数据的完整性、精度、均匀性。点通的管理是非常严格和有特点的。Johnson & Johnson在选择服务商时，我们是遥遥领先的第一，就是一个证明。
Datum has developed a unique and ISO9001 2000 certified quality-control method, which enables us to guarantee the integrity, accuracy and uniformity of clients' data. Our superior quality-control process is the primary reason for our successful bid as Johnson & Johnson's service provider.

点通有自己独创的新技术
We have implemented the following new technologies:

- 新技术之一：低质量图像的识别技术
  The recognition technology for degraded images

- 新技术之二：索引技术
  Indexing technology

  对于一般性的关键词(Key Words)，建立通用的库，用户随时可以查阅这些库。对于特定的术语，建立链接关系。
  Establish universal database for reference of general key words-- Clients can access the database anytime; Establish linkage among specified glossaries.

- 新技术之三：单一平台多语言, 多个数字资料库检索
  Information retrieval from multi-database with multi-language cross-reference on a single platform

点通正在加工中美合作的百万册图书项目。中国将在明年完成100万册图书的数字化加工，并在若干高校启用。
Datum has participated in the China-US joint Million Book Digital Library Project. This project will be completed next year. The Digital Library will be available in many universities.

点通可以发挥中国劳动力的成本优势，为全世界各个图书馆服务。这是一种真正的电子商务。在数据加工上，点通与你的邻居没有什么不同。点通公司可以通过网上传输的方式，为各个图书馆提供优质的服务。
Combining our technical expertise, the low-cost of Chinese labor force, and the convenience of high-speed Internet access, Datum can provide excellent digitization services to libraries worldwide. This is a true E-commerce.

Thank you for listening. I am happy to take any questions.

Http://www.datum.com.cn/
Http://www.datumdata.com/

E-mail: Zhangyuzhi@datum.comcn or byang@vip.sina.com