用于数字化处理及 OCR 的缩微处理

对于报纸缩微处理手册的补充

导言

由国际图联（IFLA）1996 年编辑出版的*《报纸缩微处理手册》*已经在国际图联的网站上发表。

国际图联报纸委员会（它的前身为：国际图联报纸协商会议）为了使未来对缩微胶卷数字化或进行文字识别，计划对《报纸缩微处理手册》进行补充。近期从以下几个方面的推动了这项工作的进展。

- 因特网的迅猛发展，因特网为用户查询不同年代的报纸提供了方便
- 计算机硬件性能大大提高，应用软件的性能同时提升，使得客户能从数字形式快速准确地获取文字信息
- 数字介质的存储成本降低，

年复一年，技术的发展逐步改变了人们的观念，国际图联报纸委员会考虑出版《用于数字化处理及 OCR 的缩微处理》这本手册的时机已经成熟。缩微技术在图书馆界用于保存文本资料已经应用多年。鉴于大量的报纸资料已经拍摄成为缩微胶卷，某些报纸正在进行缩微处理，优化检索缩微胶卷的手段是利用这个庞大的缩微胶卷库最合适的途径。

数字化加上有效的文字检索方法已经为方便地利用报纸提供了可能。由于图书馆报纸的利用率高，报纸的数字化已经成为图书馆数字化项目的重要组成部分。数字化对缩微胶卷的质量提出了特殊的要

求。

以缩微胶卷作为中间媒介进行报纸数字化对缩微胶卷的质量提出更高要求。在由芬兰赫尔辛基大学图书馆与瑞典皇家图书馆合作开发的北欧 TIDEN 项目已经对缩微胶卷的质量与数字化及 OCR 的效果进行了试验。因为大家对上述试验的结果十分关注，上述项目的参与单位（包括：上面已经提到的两个图书馆，还有挪威国家图书馆及丹麦 Aarthus 国家图书馆）把结果递交给国际图联报纸委员会供起草《用于数字化处理及 OCR 的缩微处理》使用。北欧 TIDEN 项目主要利用缩微胶卷作为中间媒介对 1640 年到 1900 年期间的北欧报纸进行数字化，而这个阶段的北欧报纸主要以歌德字体为主，也有使用罗马字体。这里特别还要提到国际图联报纸委员会成员法国国家图书馆与大英图书馆的合作。本文将作为由国际图联报纸委员会已经发表的《报纸缩微处理手册》的补充。

本文涉及范围

存储在电子介质的数据便于访问，缩微胶卷保存寿命长，对文献的保存及利用，这两者结合是最佳方案。缩微胶卷的保存期限可达 500 年，它可以安全地保护文字档案，对于已经发脆的纸张也可以进行缩微拍摄处理。数字化可以提供方便的存取手段。而缩微胶卷的使用比以纸为介质的档案更为便利。如果在缩微化的过程中考虑这方面的因素，以缩微胶卷作为数字化的中间介质将会更方便、经济。

本手册的目的在于：如何利用缩微胶卷作为未来数字化的平台并在已经数字化的文献中如何实现全文检索。由于数字化的技术发展速度很快，本手册着重根据目前的实际在国际图联的《用于数字化及OCR 的缩微保存处理》指导提出建议。随了技术的发展，新的方法会不断出现。缩微处理与数字化处理可以通过多种方式结合。可以先缩微处理，然后再由缩微胶卷实现数字化。也可以先进行数字化，然后利用 COM 方法（由计算机向缩微胶卷输出）将文献内容保存到缩微胶卷上。还可以利用缩微与数字化合成拍摄机同步进行缩微和数字化处理。对于报纸，一般首选缩微处理，这是最常用的方法，这里将讨论具体的细节。

必须强调缩微胶卷的拍摄质量对于数字化的图象质量是非常重要的。对于数字化后的报纸进行全文检索有赖于高质量的 OCR 处理，而它必须依靠高精度的图象。换而言之，输出取决于输入。

对于提高缩微质量的一些建议

图书馆必须把基于高质量要求的缩微处理项目作为图书馆保存文献和进行数字化处理政策的组成部分。质量要求水平取决于图书馆的需求。其中最重要的是必须依据缩微处理的技术标准，如：国际图联推荐的《报纸缩微处理手册》中所列举的标准。如果图书馆在缩微项目后，希望进一步实现自动处理数字化和全文检索，这更有必要。

缩微胶卷的特征

对于原件的文字较灰或纸质发脆、纸色返黄，高对比度的缩微胶卷可以获得较好的扫描效果。文字可以与底版明显地区分开来。可以把原件上较小的斑点、污点或皱折去除。这样便于 OCR 软件对数字图象进行识别，最终可以提供全文检索功能。缩微处理缩微比例为 16 倍时，可以较好地满足质量指标要求，当缩微比例超过 20 时，可能就无法满足质量指标的要求。

然而缩微照相不可能提供与根据原件数字化相同质量的图象。当从印刷在报纸上原始图象数字化…缩微不同种类拷贝缩微胶卷之间也有明显的区别，

由北欧 TIDEN 项目得到的结论

在由赫尔辛基大学图书馆与瑞典皇家图书馆实施的北欧 TIDEN 项目—北欧报纸数字化项目中，对从 19 世纪起到 20 世纪早期报纸，利用 35 毫米缩微胶卷对其数字化并进行 OCR 处理，比较处理结果。我们发现四个因素，与能够顺利地实施 OCR 识别或特别不能进行 OCR 识别相关。

- 报纸原件的文字质量，报纸版面布局的复杂程度及纸质
- 拍摄机的缩微比例
- 报纸文字字体和字体的大小（罗马字体识别的效果比较好，而旧歌德字体需要专门的技巧。包括歌德字体在内的多种字体常常会识别错误）
- 语言（瑞典语的识别效果比芬兰语好。多种语言混合在一起比单种语言难以识别）

拍摄机品牌、分辨率以及缩微胶卷（第一代母片还是第二代拷贝片）的影响与 OCR 识别相关。必须注意在 50 年代和 60 年代拍摄的老缩微胶卷可能无法进行 OCR 识别，而 80 年代以后根据国际标准拍摄的缩微胶卷 OCR 识别的效果较好。

优化缩微胶卷扫描过程

不论从报纸的原件进行扫描还是通过报纸的缩微胶卷进行扫描，经过文字识别后都有可能获得相同高质量的全文检索效果。但是在实际工作中，缩微处理的过程不可能达到在实验室中处理相同的水准，其文字识别效果会有差异。因此在缩微处理的整个过程中，所有的标准应该改进。

1. 不论是缩微处理还是数字化，报纸原件的质量是至关重要的，这也是为什麽要选择尽可能好的报纸作为处理原件的理由。纸质很差的报纸应该由专职保管人员对其进行处理。被处理的报纸应当是未装订。装订的报纸应当放在专用用玻璃压实的书支架进行拍摄（在这种情况下页面文字边缘与书脊之间的空白必须保持 1.5 厘米）。
2. 缩微胶卷扫描仪可能会遇到鉴别胶片上不同大小报纸页面问题。这由于缩微胶卷扫描通常根据胶片的左边或边缘来鉴别新的一帧。在这种情况下可以通过下面办法加以识别：在拍摄相机工作台拍摄区域的左边放一纸带用以标识。纸带可以进一步扩展为可供机器阅读的条码，条码中可以保存有关报纸的卷期、页、版本、增刊及重

复页等信息。这些可以利用计算机软件加以实现。

3. 用于数字化的缩微处理在制作过程中，均匀的光照十分重要，在一卷胶卷中其每帧的光照变化必须控制在 0.2 以内。缩微胶卷上每帧图象状况发生任何变化，都必须由扫描仪的操作人员加以调整，否则 OCR 的质量无法保证。每帧图象的曝光量发生变化需要对扫描参数进行人工调整，这对于连续自动胶片扫描来讲，工效低，成本会提高。

4. 对于老的缩微胶卷其帧与帧之间的距离太小，缩微胶卷扫描仪对其难以区分。在扫描时将规定相同大小的扫描区域。

总而言之，为用于保存进行缩微胶卷采用缩微处理的国际、国家标准是以后进行数字化成功的重要保证。采用未装订的报纸进行缩微拍摄以及缩微拍摄的缩微比例，对于后续的 OCR 识别处理的质量是非常重要的影响因素。为能够进行缩微胶卷自动扫描处理，图象的曝光量的变化必须控制在一个很小的范围之内。在每帧图象的开头必须用白色的带或条码标识。用来表示报纸的卷期（页码）及缩微胶卷等内容的条码可为今后将 OCR 识别的文字信息用数据库加以管理提供方便。利用计算机软件处理缩微胶卷的扫描及建立索引的过程。

财务分析

为执行扫描处理，报纸出版物的准备及扫描处理的本身在开始时要进行财务预算。这需要由图书馆与扫描代理商共同进行计划。缩微项目是以后数字化的基础，要在缩微项目开始时就要策划。

项目可能需要较大的财务预算，例如：由于降低缩微比例、将已经装订好的报纸拆开等。假如需要使用条码，缩微拍摄操作人员还必须对条码给予注意，这样也会降低拍摄速度。对于 OCR 识别而言，缩微胶卷拍摄时解象密度提高，软件的识别率可以提高，但这样有可能导致需重新拍摄缩微胶卷。对于已经装订成册的报纸，进行拍摄需增加额外的设备。上述一些内容在不少图书馆中已经付褚实施，然而这些新的要求也可以结合以后新项目予以实施。尽管开始投入的费用较高，但可以得到高质量的缩微胶卷，在随后的数字化过程中可以使它便于被机器阅读，以提高效率、降低成本。由于技术的发展，用户需要高质量的缩微胶卷进行数字化，从而最终可以不需要报纸的原件，即使报纸的原件不存在了，读者仍然可以进行阅读。这样最终实现节约开支，并一劳永逸。

技术确实会去掉我们目前所遇到的部分障碍，而不是全部障碍。如果缩微处理的质量保证程序作为图书馆总体策略的一部分的话，它将为今后的数字化过程提供方便。缩微处理应当作为图书馆数字化进程的一个组成部分。

2002 年 12 月