
IFLA

International Federation of Library Associations and Institutions
Fédération Internationale des Associations de Bibliothécaires et des Bibliothèques
Internationaler Verband der bibliothekarischen Vereine und Institutionen
Международная Федерация Библиотечных Ассоциаций и Учреждений
Federación Internacional de Asociaciones de Bibliotecarios y Bibliotecas
国际图书馆协会与机构联合会
الاتحاد الدولي لجمعيات ومؤسسات المكتبات

About IFLA **www.ifla.org**

IFLA (The International Federation of Library Associations and Institutions) is the leading international body representing the interests of library and information services and their users. It is the global voice of the library and information profession.

IFLA provides information specialists throughout the world with a forum for exchanging ideas and promoting international cooperation, research, and development in all fields of library activity and information service. IFLA is one of the means through which libraries, information centres, and information professionals worldwide can formulate their goals, exert their influence as a group, protect their interests, and find solutions to global problems.

IFLA's aims, objectives, and professional programme can only be fulfilled with the cooperation and active involvement of its members and affiliates. Currently, approximately 1,600 associations, institutions and individuals, from widely divergent cultural back-grounds, are working together to further the goals of the Federation and to promote librarianship on a global level. Through its formal membership, IFLA directly or indirectly represents some 500,000 library and information professionals worldwide.

IFLA pursues its aims through a variety of channels, including the publication of a major journal, as well as guidelines, reports and monographs on a wide range of topics. IFLA organizes workshops and seminars around the world to enhance professional practice and increase awareness of the growing importance of libraries in the digital age. All this is done in collaboration with a number of other non-governmental organizations, funding bodies and international agencies such as UNESCO and WIPO. IFLANET, the Federation's website, is a prime source of information about IFLA, its policies and activities: www.ifla.org

Library and information professionals gather annually at the IFLA World Library and Information Congress, held in August each year in cities around the world.

IFLA was founded in Edinburgh, Scotland, in 1927 at an international conference of national library directors. IFLA was registered in the Netherlands in 1971. The Koninklijke Bibliotheek (Royal Library), the national library of the Netherlands, in The Hague, generously provides the facilities for our headquarters. Regional offices are located in Rio de Janeiro, Brazil; Pretoria, South Africa; and Singapore.

IFLA Publications 150

Newspapers

Legal Deposit and Research in the Digital Era

Edited by
Hartmut Walravens

De Gruyter Saur

IFLA Publications
edited by Sjoerd Koopman

ISBN 978-3-11-025325-2
e-ISBN 978-3-11-025531-7
ISSN 0344-6891

Bibliographic information published by the Deutsche Nationalibliothek

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.d-nb.de.

Walter de Gruyter GmbH & Co. KG, Berlin

∞ Printed on permanent paper
The paper used in this publication meets the minimum requirements of American National Standard – Permanence of Paper for Publications and Documents in Libraries and Archives ANSI/NISO Z39.48-1992 (R1997)

Printing and binding: Strauss GmbH, Mörlenbach

Printed in Germany

www.degruyter.com

CONTENTS

PREFACE

2009 was a particularly active year for the IFLA Newspapers Section, and therefore the present volume offers the proceedings of more than one conference.

The spring conference took place in Mozhaisk, Russia, by invitation of the Russian Book Chamber (Moscow), the national bibliographic centre of the Russian Federation. The Newspapers Section was very pleased to be able to address an audience from Russia and former Soviet republics as there is a vast amount of newspapers waiting to be made available to a larger circle of researchers. Colleagues were very eager to profit from the experiences of foreign libraries and institutions, as there were a number of efforts already under way to preserve the newspaper holdings and facilitate access to the collections. The theme of the conference was «Legal deposit of newspapers for libraries: challenges of the digital environment», a subject which at first sounds very traditional but which comprises such hot topics as legal deposit of born digital newspapers, publishers pdf files of printed papers, and preservation and conservation of such files.

It is within the general IFLA policy that the section decided to publish the papers both in Russian and in English in order to make them available and accessible to a larger professional audience in Russia and the former socialist countries. As there was no funding for professional translations, the authors and the Russian Book Chamber worked on the English versions which leave room for stylistic improvement. The editor made every effort to see to it that the texts provide correct information and are coherent even if other duties prevented him from a more thorough rephrasing. In any case, there is a lot of new information in the contributions not easily found anywhere else.

The contributions to a satellite meeting hosted by the National Library of Sweden are also included in this volume; the Section felt that the challenges of born digital papers needed some in-depth discussion, and Dr. Gunnar Sahlin, the National Librarian of Sweden was kind enough to host the conference, entitled: «The present becomes the past: harvesting, archiving, presenting today's digitally produced newspapers». The term «born digital» was intentionally avoided in order not to eliminate issues connected with printed papers which are nowadays also «produced digitally» as well as historical papers made available in digitized form.

Finally, four papers presented at the World Library and Information Congress which took place in Milan, Italy are included. The theme of the panel was: «Newspapers in the Mediterranean and the evolution of the modern state». Quite naturally, the host country Italy was the centre of attention but a very interesting paper dealing with Palestinian newspapers provided a wider Mediterranean outlook.

The Milan conference took place at the time when the present editor's reponsibility as chairman of IFLA's Newspapers Section came to an end. This may be an opportunity to add a few words on the Section's work during the past years. Apologies for their personal note!

The present editor got involved in the then IFLA Round Table on newspapers at the New Delhi IFLA General Conference (1996), and soon became its chairman, for a time concurrently with the Serials Section. When he retired from this function in the autumn of 2009, the forthcoming International Newspaper Conference was being planned for spring 2010 in New Delhi – which completes a circle, in a way.

Like most newspaper librarians I became involved in the material by accident. Library schools usually avoid the subject of newspapers, and during my training I never had to face

such media professionally. Nor later on when working as a subject specialist for humanities and social sciences. When I took over the responsibility for the German Union Catalogue of Serials (*Zeitschriftendatenbank*; ZDB; http://zdb-opac.de), newspapers suddenly became an issue: There was an enthusiastic newspaper specialist among the staff, Marieluise Schillig, and she was in charge of a file of foreign newspaper holdings. Foreign papers were in demand, and so a small directory, *Standortverzeichnis ausländischer Zeitungen und Illustrierten* (SAZI) had been published. In the meantime this catalogue was outdated, and interlibrary loan applications were checked against the card file. The question arose whether it would not be an improvement to enter the data into ZDB and have the holdings updated by the respective institution. Fortunately support from the German Research Association (DFG) came in, and this file and other large holdings became part of ZDB which had turned into an online shared cataloguing system with about 1.4 million serial titles, among them appr. 80,000 newspaper titles. Offshoot of these activities were several offline publications, like a guide to newspaper cataloguing, and newspaper directories, as well as regular meetings of a Newspaper Working Group (Arbeitsgemeinschaft Zeitungen) in Germany.

Serials responsibility at the Berlin State Library led to an active participation in IFLA work, especially in the Serials Section which served as the parent body to a Round Table on Newspapers. The latter was a small but very active and dedicated group. It was not subjected to the rules for regular members and officers turnover, and so the few experts (some of them already officially retired) had been working there for a longer period of time. When it was my turn to chair the Serials Committee I found myself in a difficult situation: At that time the serials colleagues were mostly interested in cataloguing (several of them were also in charge of ISSN), while the newspaper people were focused on microfilming, preservation, and the new buzz word digitisation. So I was not only formally but in practice chairing two groups, and that meant double the amount of work. Fortunately much support was provided by the secretary of the Round Table, a responsibility traditionally taken by the newspaper librarian of the British Library. I had contacted the Round Table at the New Delhi IFLA, and the then chairman, Robert Harriman (Library of Congress), encouraged me. Secretary was Geoffrey Hamilton who soon retired. Bob Harriman also stepped down soon, for professional reasons. When it came to electing a new chair person, colleagues looked at me – I was the youngest one, or had the least seniority – so it was my obligation! I feebly protested that I was not really a newspaper man. It did not help, and so newspapers became an even more important part of my activities. Fortunately, the new secretary, Geoff Smith, and, after a very brief interlude with John Byford, Edmund King, were a great support, and we built a very pleasant and efficient working relationship. This was particularly important as I also had the responsibility for the Serials Section, became secretary to the Coordinating Board, was a permanent guest at the Conference of Directors of National Libraries, and as Director of both the International ISBN and ISMN Agencies had a lot of other missions to fulfill at the IFLA General Conferences.

It had struck me from the beginning that the (originally) two hour time slot for committee meetings at the General Conferences was not sufficient for efficient work. Some colleagues were involved in other sections, too, e.g. Preservation, and could not always participate in the meetings. Others did not get funding because the General Conference suffered from the reputation of being not specialized enough – why should a library send a newspaper specialist to the conference if the newspaper part lasted just two hours? A new concept and a new strategy were necessary. So additional mid-term meetings were started, at first partly in conjunction with the annual meetings of the U.S. newspaper project. These were great opportunities to meet experienced colleagues, find out about the progress of the project – cataloguing, preservation, access. Also, one could get a first-hand impression of the work of colleagues on the spot, the collections, the routine work. This approach was

successful but not really enough. At a mid-term meeting in Ottawa at the (then) National Library of Canada, the concept was modified and a first attempt at active outreach was started: Colleagues from the Library were invited to join us, and the Round Table members gave reports on what they were doing, the newspaper situation in their respective countries, etc.

Encouraged by the positive feedback, a more ambitious idea was put into practice – to combine the mid-term meeting with an international newspaper conference. Thanks to the active cooperation of colleagues from all over the world – Cape Town – Berlin – Shanghai – Canberra – Salt Lake City – Santiago de Chile – Singapore – Moscow – this strategy worked very well.

These conferences became a success, and the Round Table attained the reputation of being a travellers' club. Most people did not realize that a successful event requires a lot of hard work, organizational skills and experienced, or at least very cooperative, partners on the other side. These conferences picked hot topics as general themes, e.g. preservation and access, digitisation, and tried to involve as many colleagues as possible. The goal was to learn about the other countries, foster the relationship with colleagues, and provide information and support to the host countries.

While conferences have the great advantage of bringing people together and promote the exchange of information, the results are very often soon forgotten – there are too many meetings, and there is so much specialized information that is constantly updated. As a bookish person, I was aware of the fact that only published information would be of lasting value and would also allow those people who could not participate in the meetings, to have access to the proceedings. So the slogan was – no conference without a (good) publication! Provided the contributions were of a satisfactory level, this was not too difficult to achieve – it just meant editing, layouting and publishing: fortunately, I had been doing that already for twenty years, and as «Mr. ISBN» I had a good working relationship with publishers. A number of volumes appeared in the IFLA Publication series, thanks to the good offices of Sjoerd Koopman, of IFLA HQ, and Manfred Link, of De Gruyter Saur Publishing. A special treat was the Santiago de Chile volume – all contributions were offered in English and Spanish in order to reach Latin American colleagues who usually do not have an adequate command of English. Thus, an English-Russian volume was the next goal ...

There are, of course, limits to what an IFLA body can do, especially when it is so small; projects did not make too much progress, except for the two seminal guidelines on microfilming and digitisation, prepared by Else Delaunay and Majken Bremer-Laamanen, both very active members of the Section. A project on African newspapers stagnated for lack of funding. Also the idea of having an international newspaper handbook along the lines of Marieluise Schillig's very useful German one, could not be realized – lack of time ...

An important event was the change of the Round Table into a Section. This was by no means after the gusto of the newspaper colleagues, and we tried to escape this – it meant more red tape, and a rotating membership scheme not ideal for the small newspaper group. But there was no alternative, except remaining a special interest group to be dissolved after two years. A few years later it was ruled that a section needed a minimum institutional and delegates membership. The Newspapers Section worked hard on overcoming these hurdles and is close to reaching the goal ...

So after both the secretary and the chairman of the Section stepped down in Milan, they both have a feeling that they had an excellent time in working with dedicated, pleasant colleagues and friends, and that they also helped to get the necessary attention for newspapers within the library community and their business partners as well as alerting many people in

many countries to the necessity of quick action regarding the preservation of their cultural heritage, in print and online – especially NEWSPAPERS!

Thanks are due to many colleagues all over the world, especially the members of the Section who gave their support with unfailing enthusiasm. They cannot be named all here but I cannot but appreciate the always pertinent and practical advice of Henry Snyder (formerly of the University of California, Riverside) who contributed enormously to the success story of the Section, and last not least Ed King, hard-working secretary to the Section, who did an excellent job as an organiser and a newspaper specialist. The Section enjoyed the support of IFLA's programme director, Sjoerd Koopman, and De Gruyter Saur's producer, Manfred Link, who helped to get the results of conferences and research published and distributed. The proceedings would never have seen the light of day, however, without the unfailing support and industry of Carolin Unger in Berlin, who worked tirelessly on the layouts.

The present volume took a bit more time to produce than previous volumes. It was not so easy to collect all the contributions from three different conferences, and it was imperative to have the Russian papers presented also in English translation. Special thanks are due to my colleagues at the Russian Book Chamber in Moscow, Alexander Dzhigo and Konstantin Sukhorukov who have been good friends for many years.

While the present volume was being edited the idea came up to make the contents of previous publications by the Section more easily available by means of an analytical index of contributions. It was almost surprising to see that there were by now, including the present one, ten monographs:

Proceedings of the IFLA Symposium Managing the Preservation of Periodicals and Newspapers. Bibliothèque nationale de France, Paris, 21–24 August 2000. Edited by Jennifer Budd, IFLA-PAC.
München: Saur 2002. 175 pp.
(IFLA Publications 103.)

Newspapers in international librarianship. Papers presented by the Newspapers Section at IFLA General Conferences.
München: K. G. Saur 2003. 260 pp.
(IFLA Publications 107.)

Newspapers in Central and Eastern Europe. Zeitungen in Mittel- und Osteuropa. Papers presented to the Newspaper Section at the IFLA Post Conference, Berlin 2003. Ed. by H. W. in cooperation with Marieluise Schillig.
München: Saur 2004. 251 pp.
(IFLA Publications 110.)

International newspaper librarianship for the 21st century.
München: Saur 2006. 298 pp.
(IFLA Publications 118.)

Newspapers of the world online. U.S. and international perspectives. Proceedings of conferences in Salt Lake City and Seoul, 2006.
München: K. G. Saur 2006. 195 pp.
(IFLA Publications 122.)

Gazety / newspapers. resources, processing, preservation, digitization, promotion, information. Conference proceedings Poznan, October 19–21, 2006.
Poznań: Universytet im. Adama Mickiewicza, 2006. 478 S.

Newspapers collection management: printed and digital challenges.
La gestion de colecciones de periódicos: desafíos en impresos y digitales. Proceedings of the International newspaper Conference, April 3–5, 2007.
München: Saur 2008. XI, 396 pp.
(IFLA Publication series 133.)

The impact of digital technology on contemporary and historic newspapers. Proceedings of the International Newspapers Conference, Singapore, 1–3 April 2008, and papers from the IFLA World Library and Information Congress, Québec, Canada, August, 2008.
München: Saur 2008. XII, 222 pp.
(IFLA Publications 135.)

Digital preservation and access to news and views. Conference papers – IFLA International Newspaper Conference 2010. Edited by Ramesh Gaur et al.
New Delhi: Indira Gandhi National Centre for the Arts; IFLA Newspapers Section 2010. 284 pp.

Newspapers old and new; international perspectives. Legal deposit and research in the digital era. Proceedings of events of the IFLA Newspapers Section, 2009.
München: W. de Gruyter 2011. 308 pp.

The first volume was edited by IFLA PAC but the conference was co-sponsored by the then Round Table on Newspapers, and a number of newspaper colleagues contributed papers. Not in the IFLA Publications series but nevertheless co-sponsored by the Section are the proceedings of the International Newspaper Conference that was hosted by Poznań University Library, Poland, where Dr. Artur Jazdon gave his unfailing support, and the recent (2010) proceedings of the New Delhi International Newspaper Conference, edited by Dr. Ramesh Gaur, Indira Gandhi International Centre for the Arts. The latter conferences were also very successful, and they fostered the information flow and the relationship with newspaper experts in Poland and in India. Both publications were offered to the participants of the conferences at the opening of the respective events! Because of this they have not enjoyed a wide circulation, and IFLA may want to make them available to its worldwide audience.

Berlin, October 2010 Hartmut Walravens

OPENING ADDRESS

Ladies and gentlemen, dear friends and colleagues,

I would like to warmly welcome you on behalf of the IFLA Newspapers Section and our long-time ally, the IFLA PAC, to this conference which is kindly hosted by the National Library of Sweden.

Newspapers have been very popular in the other media during the last year, for an unexpected reason – a number of well-established papers in the United States became defunct, and it is not improbable that others will follow. People have been wondering about the factors that triggered this development. Is this due to a change of reader behaviour? Is it the competition of other media, especially the internet? Is there a connection with the worldwide economic slump? Did companies stick too much to their time-honoured business model? Should they have gone online earlier?

Certainly, it was not just one factor that was responsible for this development. Online newspapers have been claimed for quite some time to be the solution to the industry's troubles. They are (allegedly) less expensive to produce, they are easily updated, distribution is easier and cheaper, and now with the Electronic Ink project finally having hit the market readers may even have a tangible electronic paper before them. On the other hand, it is well known that most online newspaper versions did not bring in any money as advertisers preferred the paper editions, and readers, too, if it was not for quick information and updates. So are we facing a change of paradigms now? And do readers want an even lower level of infotainment? It was a nicely placed April First news item that one reputed US paper had decided to be available through Twitter only ...

Besides this current development there are challenges for the librarian and archivist. At first people believed things would become easier – electronic files would be delivered by the publishers, no collation, no binding, no storage space: wonderful! Then it turned out additional legislation was required, expensive mass storage needed; there were doubts about the reliability of storage systems, frequent control of files became indispensible. Access also provided headaches – could it be provided free of charge, or under what conditions? What really belonged to the paper? What about photographs and news agency reports? Copyright became a buzz word.

Which edition version of a paper should be collected? Should any change be preserved and recorded?

That was not all – files came in in a variety of formats, and this clearly called for standards. A Hundred Flowers Movement was not desirable in an environment that yearned for interoperablilty and synergy. These challenges make a conference like this one highly necessary and desirable.

Looking at the programme I trust that we have a very interesting and fruitful event before us. This is mainly owing to the efforts of our hosts: Dr. Gunnar Sahlin, National Librarian of Sweden kindly received us when we kind of invited ourselves, and Pär Nilsson, Senior Newpaper expert, and his team who bore the burden of the organisation. They as well as our sponsors deserve our unswerving gratitude.

In closing a few words about the IFLA Newspaper Section – it is a small international group of dedicated newspaper experts that welcomes your cooperation. Its strength is its outreach: since 1992 it has published or co-published 9 books, with a tenth now in the pipeline. And

it has organized newspaper conferences all over the world to make colleagues aware of the necessity of preserving and making available newspapers – a unique historical source! You will find more information on the Section's webpage (part of www.IFLA.org) and newsletter.

Hartmut Walravens

HARVESTING OF ONLINE NEWSPAPERS AT THE NATIONAL LIBRARY OF SWEDEN

Allan Arvidson and Oskar Grenholm

The web harvesting activity at the National Library of Sweden, named «Kulturarw3», started as a project in 1996. In the beginning the effort was concentrated towards downloads of the complete Swedish web space. The first download was completed in 1997, yielding about 3 million web pages. The total amount of data saved was about 160 GByte. This project has since then been continued as a regular activity.

In 2002 the harvesting was expanded to include also online newspapers. The starting point was a list of daily newspapers with an online version maintained by the library. It included about 140 newspapers.

In the beginning the harvesting was done with a home made harvester. Later we moved to «heritrix», a harvesting software specially designed for archival purposes. Heritrix was developed and maintained by Internet Archive in collaboration with the International Internet Preservation consortium. The software was configured to make one download of all the selected newspapers web sites once a day. In order to get only the daily issue the harvester was configured to go breadth first, i.e. not to drill down in one branch but to take the web site «layer-by-layer». The maximum number of objects harvested was individually configured for each paper. The number of objects varied from 3-400 hundred for most papers up to a few thousand for some.

The harvesting process is completely automated, the operator only has to intervene if there is an error. An automated script which checks that the last download went well runs every morning and sends an alarm if the number of harvested objects falls below a configurable limit.

Before starting the harvesting a letter was sent to all the newspapers involved explaining what we were about to do. Very few reacted to the letter and most of those who did were positive to it.

The number of objects harvested every day is about 3 Gbyte comprising about 75000 objects and in general it runs very smoothly. There is seldom reason for the operator to intervene.

There are however some problems. After starting to harvest every day a couple of papers blocked the harvester. Also, sometimes there is a problem to get a complete download from some sites. This is mainly due to extensive use of javascript. The software is very good at following static html-links. However, when a link comes from a javascript the software often fails. To treat such material correctly the software should run the javascript. This is not yet technically possible. Instead it tries to find links using regular expressions, e.g. search for things starting with «http://». This fails if the link is constructed piece by pice using e.g. browser variables. This is frequently the case with style sheets. The result of this is that although we get all the information, what's shown on the screen doesn't look like live web.

The access to the papers is via a program which allows the user to select the newspaper he or she wants and the day. He can then browse in the normal fashion. Much like the Internet Archives Wayback machine. At present access to the material is only allowed on our own premises, i.e. the user has to come to our building in Stockholm.

Many newspapers update their online version continuously. This means that the concept of a defined issue is not applicable. Furthermore, as the harvester runs once a day, it captures the state of the website at that particular point in time and all the changes in between the runs of the harvester are lost.

A quick look at the online papers reveals that most material in the web version of the newspaper is also found in the printed version. I.e. the web newspaper is not «another paper» but merely the online version of the printed paper.

As mentioned above the starting point was a list of daily newspapers that also had a web version. There are several online papers which have no corresponding printed version. Also, there are sites on the web which serve the same function as a daily paper, e.g. news blogs. This activity should be expanded to include this kind of web servers. In this work we've concentrated on daily papers. Of course there are also weekly, monthly and papers with other frequencies of issue. The web has a broad range of services that could be considered for this kind of harvesting. It is after all a task which is small enough to be managed by humans.

NEWSPAPERS AS NEW MEDIA

Pelle Snickars

A conference entitled, «The Present Becomes the Past – harvesting, archiving and presenting today's digitally produced newspapers» could hardly be arranged at a better moment. As you all know, one of the major discussions taking place in the news in general, and at various tech magazines, sites and blogs in particular during this summer has centered around and focused on the issue of the future of newspapers. In a brave new digital world, how are newspapers going to survive; in which ways do they need to upgrade themselves, how is the digital platform actually going to be used – and most importantly: how to make money online? Charging for digital content remains difficult, and online advertisement only ads up to approximately a fifth of print ad revenue. In general, newspapers have an 80-20 relation between ad incomes from print and web.

Now, various media tie-ins, on or off line, are of course one way to increase income. During this summer hardly a Swede has missed that *Aftonbladet*, the tabloid with the highest circulation in Scandinavia, is selling DVDs with all five previous Harry Potter films as a tie-in taking advantage of the buzz around the new feature film, *Harry Potter and the half-blood Prince* shown at numerous cinemas. We actually have a six-year-old Harry-wannabe at home, and the previous four Saturday evenings have all been devoted to Mr. Potter – one more to go.

Anyway, tie-ins are one way for newspapers to make money, but how to monetize online? Well, the June issue of *Wired* this summer, that is, the new U.K. version of this branded magazine, for example featured an interesting article entitled, «Can Murdoch Save Online News?». It was written by James Silver and as he noted: «It's hardly a hot scoop [that] newspapers are in deep trouble. If they survive, albeit in digital form only, then one event in May this year will surely be seen as a turning point in this narrative.» Silver was referring to a press event in May where Rupert Murdoch, the chief Executive Officer and Founder of News Corporation – one of the world's largest media conglomerates – was discussing the third-quarter results of his corporation. Apparently, News Corp's operating income had plunged by 47 percent with its newspapers division hardest hit. Evaporating advertising, meant that operating income in the news segment, including for example *The Wall Street Journal* and the *New York Post* in the US, and *The Sun* and *The Times* in the UK, were down dramatically. But strangely, Silver noted, «given this flurry of bad results, Rupert Murdoch was in a surprisingly upbeat mood». That it is possible to charge for content on the web, is obvious from the *The Wall Street Journal's* experience, Murdoch stated. Visitor numbers at wsj.com in fact doubled from 13.4 million in April 2008. So at the moment, Murdoch confessed, «we are in the midst of an epochal debate over the value of content and it's clear that, for many newspapers, the current model is malfunctioning.» And furthermore: «You can [all be confident that] News Corp. are leading the way in finding a model that maximizes revenues and returns … [we are trying hard] in devising clever ways to monetize the content of some of our long established print properties.»

Even if a media tycoon as Robert Murdoch seems confident that his corporation will find ways to monetize online, doubts remain that anyone used to a decade of free digital content will pay for news – particularly if they can get it free elsewhere. For specific content it might be possible to carve out a niche – like *The Wall Street Journal* with its solid subscription model, for instance – but charging for more general content will remain difficult.

Now, what I would like to do in this presentation is to give a brief panoramic overview of the current digital media landscape with an emphasis on newspapers as a converging new media form. In the following I will try to situate online newspapers within a broader binary media scape, and in fact only say a few, brief words about selection and archiving of newspapers. So, this is all likely going to be more of a media scholar's perspective rather than a librarian's – but bear with me. Yet since our IT-specialist Allan Arvidsson will make a presentation tomorrow of this national library's harvesting strategies of online newspapers, you will get to know more of what we actually do.

As you all know the world wide web has gone through numerous changes since the early 1990s; recently we have witnessed the so-called web 2.0 transition, and perhaps it is soon time for the semantic web – who knows. I believe, however, that the 2.0 revolutions of social network sites and the hybrid character of the web, hovering between community and commerce, are still poorly understood, not the least within the library sector. According to Tim O'Reilly web 2.0 is «the business revolution in the computer industry caused by the move to the internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: build applications that harness network effects [which get] better the more people use them.» in short, the upgrading of the web can be described as a shift from websites with static information to new sites working more as interlinked, dynamic computing platforms.

New digital media, then, functions different from traditional media. New media has, for instance, replaced the «one-to-many» broadcasting model of traditional communication with the possibility of web based «many-to-many» communication. In fact, the very foundation of web 2.0 is based on the latter model – or put more precisely «many-to-few» communication. Web pages, blogs and social networking sites are media forms functioning according to a logic where information and announcements are communicated by numerous people, but often only noticed by a few. However, new media is also said to be distinguished by its interactivity and its networkable nature. Because of its binary character it is also regularly described as being manipulable. There are for instance numerous so-called «Photoshop makeovers» on YouTube where a photograph is digitally re-edited into an image completely different from the original.

Anyway, shifting from the characteristics of new media to newspapers, the present condition of the news has of course been widely debated as the industry has «faced down soaring newsprint prices, slumping ad sales, the loss of classified advertising and drops in circulation», as Wikipedia informs us in the entry «Future of Newspapers». Revenue has, hence, plunged while competition from internet media has squeezed older print publishers. As you are all aware of, online strategies have varied over the last decade. For a while *The New York Times* for instance, tried to charge for some of its content online, attracting around 230.000 paying customers worldwide. But in 2007, reflecting a growing view in the industry that subscription fees were not able to outweigh the potential ad revenue from increased traffic on a free site – on the web money follows users – the paper announced that it would stop charging for content. With over 20 million unique visitors in March 2009 *The New York Times* has the most visited newspaper site on the web.

However, the decision to offer free online content with digital advertising as the only revenue stream, now seems to have come back to bite the industry hard. In the before mentioned article in *Wired*, the *The Financial Time*'s chief business commentator, John Gapper for example stated that: «There was an awful lot of nonsense talked about the desirability of making everything free on the net. But it was never a business strategy, only a slogan.» And a week ago, John Ridding, the chief executive of *The Financial Times* asserted that it «has become pretty clear that advertising alone is not going to sustain online business mod-

els. Quality journalism has to be paid for.» The problem at the moment is that the current recession has led to web revenues leveling off. According to analysts, ad revenue is probably down at least 20 percent compared with last year. Hence, the greater question remains as to whether new technology has in fact rendered newspapers in their traditional form obsolete.

Nevertheless, the current media transition from print to web of course has its historical predecessors. In fact, newspapers have gone through many critical media transformations, always trying to be new and modern in terms of form and content. The film clip you are now watching, is for instance, taken from a joint film venture between *Stockholms-Tidningen* and Swedish Film industry, in an effort to depict the everyday life on a newspaper in Sweden in the mid 1930s. One medium is here presented through another, as the viewer is shown various editors and journalists working for the sports or the art section. In general, the film gives a vivid impression of a distinct media modernity; a world already networked through telephone and telegraph wires shaping the news business as well as new consumer patterns.

Now, since I am originally a film scholar, and the National Library of Sweden now has audiovisual collections as part of its archive, I cannot help but show you yet another press historical film clip. This one is about how to run a newspaper – and purposely keep on loosing money.

As most of you know Orson Welles' *Citizen Kane* is the story of the rise and fall of Charles Foster Kane, a fictional character based on the newspaper magnate William Randolph Hearst. The film – often said to be the best ever made – traces the life and career of Kane, ingeniously played by Welles himself, a man whose career in the publishing world is born of idealistic social service, but gradually evolves into a ruthless pursuit of power.

If Welles portrayed the news business as a «philantropical enterprise» in 1941, more than sixty years later the most successful online newspaper is still loosing money, and chairman Arthur Sulzberger of *New York Times* is reported to be working on a new online financial strategy. On the one hand it means delivering the *Times* on different platforms, which the co-operation with Amazon and its Kindle DX is one example of. For less than ten dollars a month the *New York Times* can be downloaded to the Kindle. Since Stig Nordqvist tomorrow will talk about mobile digital e-reading I won't go into that now. Let me just say that if rumours are true that Apple are working on a media iPad – this will mean fierce competition for Amazon. On the other hand, the online financial strategy of *New York Times* will in all likelihood use some kind of «metering», that is, charging readers after a certain word count or number of clicks. An advantage of the web is that everything can be tracked and traced.

Of course, doing business online is different, but looking at the short history of the web sometimes gives one clues to the current state of things. In Sweden for example, *Aftonbladet* established itself as the first newspaper online already in August 1994. Ever since it is one of the most popular Swedish sites, with traffic approaching three million unique users every week out of a population of nine million. Most established newspapers actually went online in the mid 90s at a time when the web became more and more popular. The mid 90s was also a period when a 28.8 dial-up modem made users stare at a blank screen for several minutes. It is hard to remember today, but web browsers prior to Netscape Navigator were not able to show anything at all before all information had been loaded – and by the way, take a look at the chart of dropping usage of the Navigator browser. Anyway, the 1993 Sun Microsystems's slogan, «The Network Is the Computer», was seen by many as a true insult. Yet, only three years later two young PhD students at Stanford

began to perceive the web's network of networks of computers and servers as one gigantic information processing machine, which through sophisticated communication protocols could share bits of data and strings of code – and where «search» became the most important feature online.

So, for newspapers to thrive on the web connection speed has been crucial – and it is speed that has led to the well-known situation where a web page of a newspaper cannot be separated and distinguished from a web page of a TV channel. The difference between dn.se and svt.se is arguably a difference of degree rather than a specific difference. Interestingly, even though Swedish Television has been very successful in upgrading itself to the digital platform, it remains outranked by the tabloid *Aftonbladet* web-TV service in terms of unique users.

One way of addressing online newspapers is to compare them with new media proper, that is comparing the structure of old media migrating into new media platforms, with online startups and new media forms without an analogue past. In what ways are online newspapers for instance different from say, YouTube? My example is, however, not chosen randomly; YouTube has been the fastest growing site in the history of the web – and I have actually recently edited a book entitled, *The YouTube Reader* published by this library. So, let me take the opportunity of promoting my book by way of a slight comparison.

One of the peculiarities of YouTube lies in the way this so called media platform has been negotiating and navigating between community and commerce. If YouTube is anything, it is both industry and user driven. Online newspapers, of course, try to promote interactivity, foremost by enabling comments on articles – which sometimes are extremely interesting. Personally, I am a keen reader of *The New York Times* tech pages online, and comments there are quite often as good as articles themselves. Like the blog phenomena, comments are the typical way of distinguishing between static and dynamic text; in short, the article is not «finished» until comments seize to be uploaded. From an archival perspective this is naturally a problem. If, on the one hand, comments are left out in the harvesting process of online newspapers a defining aspect of them is lost. But the dynamics of comments, on the other hand, makes it hard to know when an article has «ended». Basically, it is the same conceptual storage problem as to when an online newspaper actually exists, since it is constantly being updated – something that the RSS-technology gives a vivid impression of. An RSS document, that is a feed or a web feed often includes full or summarized text plus some metadata – and the screen shot here is taken from the Tech page of *New York Times*. Feeds benefit publishers by letting them syndicate content automatically, and they benefit readers who want to subscribe to timely updates. But from an archival perspective they are of course extremely problematic.

Now, numerous media researchers have analyzed the interactive web, stressing its creative and grassroots potentials. Still, interactivity always has its limits. According to the so-called «90-9-1 rule» – 90 percent of online audiences never interact, nine percent interact only occasionally, and one percent does most interacting. Hence, ordinary users of a news site for example, hardly see themselves as part of a reading community. In order to facilitate such an atmosphere, media corporations have to think more like a site as YouTube. In his new book, *Remix – Making Art and Commerce Thrive in the Hybrid Economy* Lawrence Lessig for example, makes the claim that for online commercial ventures it remains crucial to transform themselves into the sort of «hybrid economies» that has come to characterize the web. YouTube for instance, owned by Google as you all know, on the one hand presents and views itself as a platform, and not a regular media distributor – especially when copyright issues are involved. Yet, on the other hand, Google clearly is a vertically integrated corporation operating in distributed ways. Bits of Google are all over the web, and both the

migration of videos on YouTube to new and old media and the embedding of clips into various sites, blogs and social-networking platforms is undoubtedly crucial for understanding the success of the site. Like Google, YouTube has distributed itself constantly. Whereas youtube.com rapidly established itself as the default site for online video, with average users and dedicated partners using the platform to perform their interests, the public also encountered YouTube videos everywhere on and off the Net. YouTube thus was and is both a node and a network.

So, according to Lessig and others – as for example Chris Anderson in his new book Free – what online newspapers really need to do to monetize content is to learn from social networking sites or so called freemium web services where basic content is free. In short, freemium is a business model that works by offering basic services for free, while charging a premium for advanced or special features. The ingenious «app store» for iPhone and iPod developed by Apple works in this way; customers can often download a free limited version of an app – and if they like it they can buy the full version etcetera. It does remain essential to price discriminate, which the app store is a great example of. Personally, I believe that the app store with its more than 1.5 billion downloads is perhaps the best example of that it is possible to charge for binary code. As trust and reliability of online technology increases, and where computers and phones become important parts, not only of work but of life itself, people will spend money on digital content. Getting people to pay will, however, require new strategies. Or as a commentator on the previously discussed *Wired* article on Murdoch stated: «If you want people to pay for [news] you're going to have to make it tailored and put them in control.» And, interestingly he added: «A Last.fm for news.»

Last.fm is a UK-based radio and music community website which claims some 30 million active users. Another, similar site is the Swedish application Spotify, a media darling and a proprietary P2P streaming service that allows instant listening to specific tracks or albums with almost no buffering delay – it just works fantastic! Nevertheless, two things are interesting and worth stressing here. The first is the very mentioning of a music site in relation to newspapers. Yet, this is what media convergence is all about. In binary form media now works in the same way, and the differences between text, image, sound and video is slowly disappearing – and so are the business models for these various media forms. The second thing is that Spotify, for example, works according to the freemium logic. Everyone has access to the site and its four million tracks; just download the application and listen. The interface features advertisement and every fifth song or so is interrupted by ad jingles. However, if one buys the premium model there are no ads whatsoever. Still, even as Spotify now has more than two million regular users, only some 40.000 have bought the premium access, even though it its only ten euros a month. It remains to be seen if the long awaited Spotify app for the iPhone, which only premium customers will be able to download, will increase the amount of paying users.

Of all the «old» media, *The Economist* lamented in 2006, «newspapers have the most to lose from the internet». A year later, Chris Anderson published a blog post where he stated that «the future of media is to stop boring us with news that doesn't relate to our lives.» So in order to sum up, what are the major challenges facing the news business? Well, let me return to Rupert Murdoch once again. Last year he gave an interesting presentation at Georgetown University – partly accessible through fora.tv – where he explicitly referred to what he called «creative destruction», a situation where new technology is constantly tearing down old ways of doing business. Let's look and listen to what the apparently 132nd richest person in the world has to say about the future of media.

Now, you can indeed say many things about old Rupert, but his analysis in this talk isn't bad at all. But what about the specific future of newspapers? Well, again I think that in order to understand the news business one has to look at other media as well. Online, there are a few persons who, even though they are sometimes way to talkative and act as opinion machines, seem to have a genuine feeling and notion of where things are heading in the binary world. Nicholas Carr is one of them, and another blog which I try to follow as often as I remember, is Jeff Jarvis «Buzzmachine». Jarvis is an egocentric media critic, who recently published a book on Google, which in fact is not bad at all. In a clip with him that I found recently, he argues that newspaper should abolish print and go fully online. Jarvis talks about a future «ecosystem of news» where the web's networks of networks produces content in various forms. Let's watch a final clip and listen to what he has to say.

Jeff Jarvis might be wrong in his analysis of the future. Personally, I think his understanding of the web, however, is way more subtle than, say, the current dispute between Google and news organizations trying to implement the so called ACAP, the Automated Content Access Protocol, a method providing machine-readable permissions for content. One of ACAP's initial goals has been to provide rules to search engine crawlers when accessing websites. In short, ACAP would prevent say, Google news from accessing content, but Google has refused to make use of the technology. In Sweden, the newsgroup Stampen, with some twenty newspapers under its umbrella, recently implemented the ACAP protocoll. Still, Google is of course driving the major bulk of web traffic to these newspapers – so why complain. In many ways the discourse around ACAP mimics the music industry's battle with file sharing since the project focuses on the needs of publishers, rather than readers. On the distributed web, the wealth of networks, to paraphrase Yochai Benkler's great book, will definitively not be achieved by cutting distribution.

STANDARDS FROM THE OTHER SIDE:

AN OVERVIEW OF THE NEWS INDUSTRY TEXT FORMAT (NITF) AND ITS KIN

Frederick Zarndt

Planman Consulting Inc.

1. Overview

The International Press Telecommunications Council IPTC (http://www.iptc.org) is a «a consortium of the world's major news agencies, news publishers, and news industry vendors. It develops and maintains technical standards for improved news exchange that are used by virtually every major news organization in the world.»[1] It is registered in London UK and as of August 2009 is comprised of about 70 companies drawn from all continents except South America.

According to Wikipedia the «News Industry Text Format (NITF) is an XML specification published by the IPTC that is designed to standardize the content and structure of individual ***text*** news articles [emphasis added]. The NITF specification defines a standard way to mark up an article's content and structure, as well as a wide variety of metadata that different organizations may choose to use.» From its beginnings in the late 1990's, NITF is today widely used by news agencies and news publishers.

This paper gives an overview of the IPTC, the NITF standard, and the related standard, NewsML. It will also suggest ways in which these standards might be used by cultural heritage institutions in programs for preservation and dissemination of both print historical newspapers and born digital newspapers.

2. Background

The IPTC «was established in 1965 by a group of news organizations ... to safeguard the telecommunications interests of the world's press.»[2] Since 1970 its focus has been primarily on the development of standards to facilitate the interchange of news. According to the IPTC management committee its objective is «to establish and maintain an open, apolitical international forum to promote and enable the exchange of news information in an efficient manner, while maintaining the highest technical quality. At the same time taking advantage of the advances in telecommunication and computing technology.»[3]

The IPTC gives several reasons for joining IPTC (here quoted from its website):

«IPTC is the only organisation that addresses the news industry concerns for standardization of information transfer formats.
IPTC is an organisation concerned with news agencies and their customers' information transfer problems.
IPTC fosters exposure to business ideas used around the world to distribute news.

1 Quoted from http://www.iptc.org About page.
2 Ibid.
3 «How To Join the IPTC». http://www.iptc.org

IPTC encourages personal relationships among peers from around the world.
IPTC provides a world news lobby voice for standardization of telecommunications services.
IPTC allows members to request research and development in areas of specific interest to their business activities.»

Membership in IPTC is as a nominating or associate member, and, as you will guess, associate members do not have same rights as nominating members. Nominating members are from organizations concerned with news collection, distribution, and publishing. Associate members may be from these organization or from vendors supporting the news industry. As of August 2009 and according to the IPTC website there are 70+ nominating and associate members. A partial list of organizations that use NITF is

Nominating members
- Agence France Press (France)
- ANSA (Italy)
- Associated Press AP (USA)
- BBC Monitoring (UK)
- Deutsche Presse-Agentur (Germany)
- Dow Jones & Company (USA)
- AB Kvälistidningen Expressen (Sweden)
- LexisNexis (USA)
- The New York Times (USA)
- Pressetext Nachrichtenagentur (Germany)
- Thomson Reuters Limited (UK)
- Tidningarnas Telegrambyra (Sweden)
- United Press International UPA (USA)
- World Association of Newspapers WAN

Associate members
- Agencia EFE (Spain)
- Athens News Agency ANA (Greece)
- AS Norsk Telegrambyra (Norway)
- BVPA (Germany)
- Canadian Press (Canada)
- CCI Europe (Denmark)
- IFRA (Germany)
- ITAR-TASS (Russia)
- Mecom (Germany)
- MENA (Egypt)
- Profium Oy (Finland)
- Ritzau Bureau I's (Denmark)

The IPTC has created several standards to facilitate the exchange of news. All IPTC standards are open, freely available, and reasonably well-documented.

- NITF is an XML based standard whose purpose is to facilitate the exchange of ***text*** news
- NewsML-G2 is an XML based standard designed for the exchange of ***multimedia*** news
- SportsML-G2 is an XML based standard for sharing sports data
- EventsML-G2 is an XML based standard for describing events
- NewsCodes are XML based metadata taxonomies for describing news items

EventsML-G2 and *SportsML-G2* are quite recent and have not yet been widely adopted.

Besides these descriptive standards IPTC has developed four metadata taxonomies for the news industry called IPTC NewsCodes. These taxonomies are

- Descriptive NewsCodes: Taxonomies to describe the content of news items
- Administrative NewsCodes: Taxonomies for administration of news items
- Transmission NewsCodes: Controlled values for the transmission of news items
- Exchange Format NewsCodes: Taxonomies to support the different news exchange format standards

NewsCodes are intended to represent concepts which can be used to categorize news content, or, more succinctly said, NewsCodes are standardized metadata about news items. «Codes have the advantage that they can be easily shared ... and, as each code requires an explicit and comprehensive definition, not only the codes but also their semantics can be shared. NewsCodes are language agnostic thus the code is the same for describing content

in different languages.»[4] Even a superficial discussion of NewsCodes requires considerable time and many more words than allocated for this paper, however, we will have more to say about Descriptive NewsCodes below as they are most applicable to text news preservation.

3. NewsML-G2

The primary focus of this paper will be NITF since it most closely aligns with XML standards commonly used in historical newspaper digitization projects. However at least passing mention must be made of *NewsML*. *NewsML* is an IPTC standard which is designed for the exchange of ***multimedia*** news in much the same way as NITF is designed for the exchange of ***text*** news. NewsML-G2 is the more sophisticated successor to NewsML and is part of the IPTC's G2 XML news standards family. NewsML-G2 is a XML-based envelope which organizes news files of almost any type. NewsML, hereinafter referring to both NewsML and NewsML-G2, is frequently used in conjunction with NITF. The latest update to NewsML-G2 (version 2.2) was released March 31, 2009.

4. News Industry Text Format

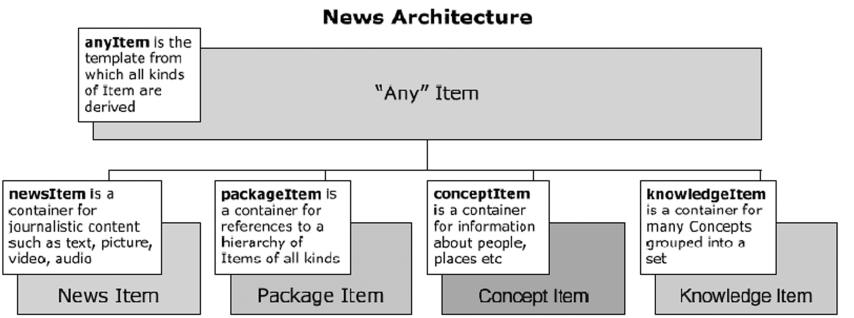

Figure 1 IPTC G2 News Architecture Framework

Standardized Generalized Markup Language SGML was used as the basis for the first version of NITF. It succeeded ANPA 1312 and IPTC 7901, both earlier standards for exchanging news content via telecommunications channels. With the introduction in 1998 of XML as an SGML subset, NITF was modified to be XML compliant. NITF version 3.4 was released May 2007 both as a XML schema and a DTD. Today NITF is the most commonly used XML vocabulary amongst news publishers worldwide and is often used with other, more recent IPTC standards such as NewsML-G2.

NITF supports the identification and description of a number of news characteristics. News articles encoded as NITF may include:

Who owns the copyright to the item, who may republish it, and who it's about.
What subjects, organizations, and events it covers.

4 Paraphrased from http://www.iptc.org

When it was reported, issued, and revised.
Where it was written, where the action took place, and where it may be released.
Why it is newsworthy, based on the editor's analysis of the metadata.

Figure 2 is a conceptual diagram of a NITF XML document. (Note that the diagram is partial: It does not illustrate all container types or fields, only the ones most interesting for this discussion.)

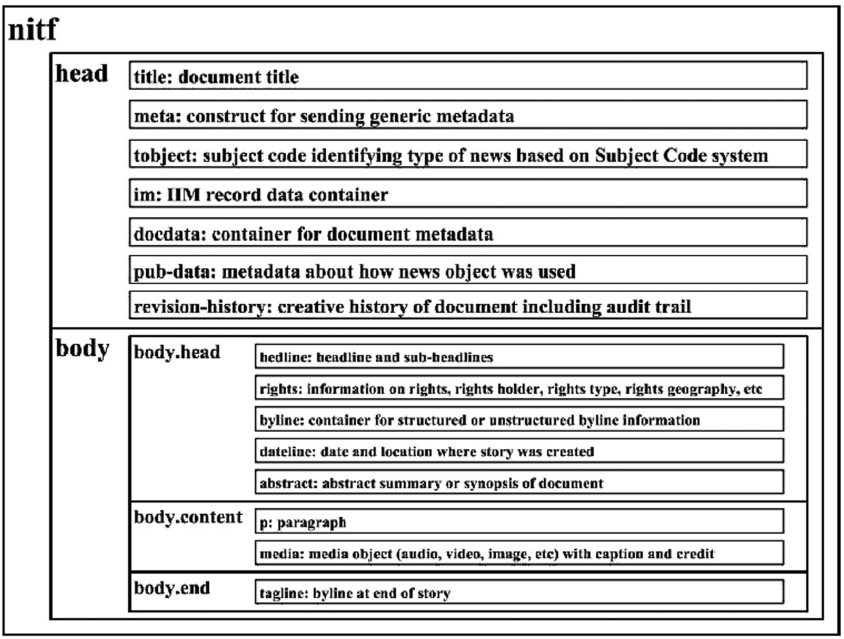

Figure 2 NITF Document Container

The optional *head* of an NITF document holds metadata about the remainder of the document and generally is not intended for direct display to the end user. All content in the mandatory body of the document is intended for display to the end user.

Neither the *nitf.head* container nor any of the elements in the *nitf.head* container are required. All elements[5] in *nitf.head* may be omitted or occur once; *meta, pub-data*, and *revision-history* may occur more than once in a single NITF document. Four elements in *nitf.head* are particularly interesting, namely, *tobject, docdata, pubdata*, and *revision-history*.

When present, *tobject* identifies the type of the news item based on a system of IPTC subject codes (more about subject codes below). A *tobject* is comprised of a *tobject.property* and a *tobject.subject*, each of which may be repeated.

5 More complete explanations and examples are given in the NITF documentation and in the IPTC Standards Guide for Implementers, both found at http://www.iptc.org

pubdata is an optional item which may contain a number of sub-items. From a librarian perspective the most interesting are *pubdata.date.publication, pubdata.name, pubdata.edition.name*, and *pubdata.issn*. each with the meanings implied by its name.

docdata is a container for a variety of document metadata all of which may be omitted, may occur once, or more than once. *docdata* may include the date that the news item was issued, when it should be released, when it should expire, etc. It may also include a statement of copyright (*doc.copyright*) or, when the rights holder is not the same as the copyright holder, a statement of rights for use of the document (*doc.rights*).

The optional *revision-history* element may be repeated and shows the reason for the revision (*revision-history.comment*), the name and job function of the person making the revision (*revision-history.name* and *revision-history.function*), and the normalized date of the revision (*revision-history.norm*). *revision-history* provides an audit trail of the changes made to a news item.

As already mentioned all content in the *body* is intended for display. The *body* consists of an optional *body.head*, a mandatory *body.content* (which may be repeated), and an optional *body.end*. The *body.content* is just what it sounds like. With version 3.4 it has 14 sub-elements each of which may be omitted or repeated. Some of the elements such as *p, hl2, table, ol*, and *ul* bear more than a passing resemblance to similar HTML markup. *body.content.media* is a generalized media object which can be used to encode a reference to image, audio, or video objects.

The optional *body.head* contains 8 constructs for headline and sub-headline (*body.head.hedline*), byline (*body.head.byline*), dateline (*body.head.dateline*), abstract (*body.head.abstract*), etc. Rights information (*body.head.rights*) may be included here instead of or in addition to the rights in *head.docdata.doc.rights* and *head.docdata.doc.copyright*.

body.end is optional and may include a tagline (*body.end.tagline*) or bibliographic information (*body.end.bibliography*).

The appendix shows an example of an NITF encoded news article rendered with a XSLT stylesheet.

5. Photo Metadata

IPTC's most widely used standard is that for photo metadata: The IPTC Photo Metadata Standard includes Core and Extension specifications. This standard is used by professional and amateur photographers alike whether for photojournalism or for some other purpose. The latest versions (Core 1.1 and Extension 1.1) of these standards were issued in July 2009. Since 2007 an annual conference has been held to discuss the exchange of visual media. Except for this brief mention, this paper will not discuss the IPTC Photo Metadata Standard, but more information about the standard can be found on the Photo Metadata page at http://www.iptc.org.

6. Descriptive NewsCodes

NewsCodes are intended to represent concepts which can be used to categorize news content, or, in librarian-speak, NewsCodes are standardized metadata about news items. «Codes have the advantage that they can be easily shared ... and, as each code requires an explicit and comprehensive definition, not only the codes but also their semantics can be

shared. NewsCodes are language agnostic thus the code is the same for describing content in different languages.»[6]

Subject Codes is a three level system for describing content by a well defined set of terms. Topics of level Subject provide a description of the editorial content of a News at a high level, a SubjectMatter provides a description at a more precise level, and SubjectDetail at a rather specific level.

Currently there are about 1400 subject code terms available and several of them can be assigned to a single news object, thus enabling a very narrow description of the content (or several descriptions).

7. Conclusions and recommendations

For libraries and other cultural heritage institutions, here's the interesting part of the IPTC standards, especially NITF and the Descriptive NewsCodes: Ways that these standards might be used within the library universe. Some suggestions

- Join IPTC
- Adopt / adapt news metadata taxonomies
- Request NITF news feeds for archival and dissemination purposes
- Transform articles from digitization programs into NITF XML

7.1 Join IPTC

Its membership page states that IPTC is a consortium of news agencies, newspaper, and news system vendors, however it does also say that small companies and sole practitioners are welcome to join. Why not libraries, too? Libraries, although not producers of news, are certainly important consumers of news and as such are (or ought to be) interested in the forms of and the means by which news is distributed.

Until the last half of the 20th century, libraries consumed news nearly exclusively through print media and have very well developed processes for preservation of printed materials including newspapers. Libraries also have well developed taxonomies for the classification of its materials and definitely have much more experience in applying them. Perhaps IPTC could benefit from library experience.

Only since the mid-1990s or so did libraries begin to focus on preservation of digital formats and on the conversion of analog print media to digital formats. And first in the 1970's did news agencies develop the primitive predecessors to today's news distribution standards such as NITF and NewsML. With libraries focused mostly on preservation and news organizations mostly on distribution seems like collaboration could lead to some synergies ...

7.2 Adopt / adapt news metadata taxonomies

Libraries have developed extensive taxonomies for classification of their materials, however, probably because of the relative immaturity of newspaper digitization programs, library-developed taxonomies for news articles are neither standardized nor uniformly applied. For example, one national library may classify articles as *News, Family Notices, Advertising*,

6 Paraphrased from http://www.iptc.org

and *Detailed Lists* while another may classify articles as *Article, Obituary, Advertisements, Illustration, Letter*, and *Miscellaneous* and yet another may choose not to classify articles at all. Adopting IPTC Descriptive NewsCodes, or a subset would give a standard, language-independent taxonomy for newspaper articles, one that is congruent with the news industry. Such a taxonomy would facilitate searches across newspaper collections and reduce or eliminate search software customization.

There are a couple of drawbacks to doing so however. First, some libraries already have adopted a taxonomy so if a standard taxonomy is adopted, a mechanism to map the existing taxonomy to the standard one may be required. Second, 1400+ descriptive codes for news articles may be overwhelming to users. Reducing or condensing the IPTC set of news codes to a smaller, more tractable set may be desirable.

7.3 Request NITF news feeds

Request NITF news feeds from news agencies? Perhaps this is a stretch and one which would certainly be opposed by LexisNexis and other similar organizations. But on the other hand, national libraries are mandated to serve as repositories of information and, in some countries, also mandated to be the institution for legal deposit. Printed newspapers are legally deposited with national libraries. Is legal deposit of born digital news not required merely because legal deposit laws have not yet caught up with the digital age?

Some countries, notably France, Singapore, and Finland, already do collect for legal deposit PDF files from which born digital newspapers are printed while others are considering such a mandate. These files require processing to render them suitable for current digital library software. Why not collect the stuff from which these files are built? NITF files require no or little further processing, only a software system to preserve and disseminate.

Even if only print PDF files are collected, these can be transformed into NITF or into METS/ALTO with greater accuracy and less cost than from image files requiring scanning and OCR.

7.4 Transform articles from digitization programs into NITF XML

Digitization of historical newspapers is most often done to article level. Since news publishers through the IPTC have provided a widely used standard for the exchange of text news articles, should cultural heritage institutions consider using NITF? There are strong reasons for doing so as well as lame reasons for continuing the METS/ALTO status quo.

NITF and related standards have been created by the news industry for exchanging news items in a standardized fashion. These standards have evolved over many years, since 1979 – longer by far than METS (2002) or ALTO (2004) have been around. They have been built to purpose for news items whereas METS is generalized for many types of content and media.

Not to suggest that NITF should replace METS or ALTO for newspaper digitization projects but it could perhaps supplement METS and ALTO. Presently articles are encoded as metadata within the METS file. No reason that NITF files could not be linked in addition to or instead of current article metadata.

Appendix: Sample NITF Article

The NITF XML code below was invented by the author to show a concrete example of NITF encoding.

```
<?xml version="1.0"?>
<!DOCTYPE nitf SYSTEM
"http://www.iptc.org/std/NITF/3.4/specification/dtd/nitf-3-4.dtd">
<nitf>
<body>
      <body.head>
            <hedline>
                  <hl1>Ed King Attends IFLA Conference in
Stockholm</hl1>
                  <hl2>A simple NITF article</hl2>
            </hedline>
            <byline>
                  By <person>Frederick Zarndt</person>
                  <byttl>IFLA Conference Reporter</byttl>
            </byline>
      </body.head>
      <body.content>
            <media style="align:left">
                  <media-reference
                        mime-type="image/jpeg"
                        source="ed-king.jpg"
                        height="185"
                        width="278"
                        >
                  </media-reference>
                  <media-caption>
                        Ed King at the IFLA conference in Stockholm.
                  </media-caption>
            </media>
            <p xml:lang="en-US">Whilst attending the IFLA Newspaper
Interest Group conference in Stockholm, Ed King was quoted as saying
"The weather is superb today!"</p>
            <p>Happy conferencing everybody!</p>
      </body.content>
      <body.end>
            <tagline>Hartmut Walravens contributed to this
article.</tagline>
      </body.end>
</body>
</nitf>
```

Here's the NITF code above rendered using a purpose-built XSLT stylesheet.

Ed King Attends IFLA Conference in Stockholm

A simple NITF article

By **Frederick Zarndt**
IFLA Conference Reporter

Whilst attending the IFLA Newspaper Interest Group conference in Stockholm, Ed King was quoted as saying "The weather is superb today! Happy conferencing everybody!"

Photo: British Library
Ed King at the IFLA conference in Stockholm.

Hartmut Walravens contributed to this article.

Another example of NITF encoding and its transformation with a stylesheet can be found at http://iptc.org/cms/site/index.html;jsessionid=aLXLl67zEnmd?channel=CH0156.

THE BRITISH LIBRARY NEWSPAPER STRATEGY: DEVELOPING COLLABORATION WITH PUBLISHERS TO DIGITISE BACK RUNS AND TO INGEST BORN DIGITAL NEWSPAPERS

Patrick Fleming

Associate Director, Operations and Services, The British Library

Abstract: The BL Newspaper Strategy aims to encompass some major developments; the move of the collections is being planned, as is the assimilation of the newspapers reading room at the BL in St Pancras, Central London. A major feature is the drive to open up the collections of newspapers, both old and new, to greater numbers. This paper will summarise the recent achievements relating to the digitisation of UK newspapers in BL collections published up to 1900. It will also focus upon recent discussions with publishers, and how the BL is working to ingest, store and to present digital copies of current newspapers published in UK regions, in BL reading rooms.

The British Library has one of the world's finest collections of newspapers.[1] Through legal deposit the Library collects an edition of most newspapers published in the UK and Ireland. It contains over 53,198 separate print titles and 370,000 reels of microfilm on nearly 50 km of shelf space. This accounts for 95% of the 1,259 titles currently available. These comprise:

- 111 national and regional daily titles
- 17 national and regional Sunday titles
- 511 paid-for weekly titles
- 637 free weekly titles

It announced its future intentions with regard to its newspaper collections in a press release in March 2007.[2] Since 2007, The British Library has put together a Strategy for its Newspaper Collections. The rationale for change is strategic, to both protect and preserve the collection for future generations and to continue to meet the ever changing digital demands of the current and future researcher. The collection contains UK titles as well as 200 overseas titles and is at risk. 15% of the collection is unusable because of its deterioration, and there are pressures on space and a need to develop greater digital content for access.[3]

The factors driving immediate change are based on the following:

- Insufficient storage
- Poor storage conditions
- Deterioration of hardcopy newspapers
- Fragmented reader experience
- Separated business processes

---

1 For further details of the collections, see: http://www.bl.uk/reshelp/findhelprestype/news/blnewscoll/index.html (visited 22.5.2009)

2 See: http://www.bl.uk/news/2007/pressrelease20070301.html (visited 22.5.2009)

3 See: Fleming, Patrick & Spence, Phil: The British Library Newspaper Collection: Long Term Storage, Preservation and Access http://liber.library.uu.nl/publish/articles/000259/article.pdf (visited 21.5.2009)

- Collections security
- Changing needs of the researcher

To meet the future demands of the reader and satisfy the preservation demands for protecting hard-copy newspapers, the British Library must change the existing storage and access model in the short to mid term. To achieve this, a number of strategic aims have been stated. What concerns us here is the aims as they relate to the creation of digital surrogates for back runs of newspapers and the ingest of born digital newspapers. The Library's plans are:

- provide the digital infrastructure to enable the collection of both digitised and born digital newspapers. Digitised newspapers will be from either existing hard copy or microfilmed newspapers. Born digital newspapers are those provided by publishers and vendors as electronic digital files
- move to digital as a preservation medium when reliable storage standards come into existence
- create a digital production facility for content ingest at Boston Spa and to prepare newspapers for direct digital delivery onto the Library's DLS in 2009

Digital technology has enabled newspaper publishers to transform production processes. All national and regional daily newspapers and an ever growing number of weekly newspapers are produced digitally with an output in PDF format. Within five years, the entire UK and Irish newspaper publishing industry will be produced using digital technology. This, together with the growth in colour presses, has enabled newspapers to increase the number of pages and products that they offer.

If the lifetime of storage for the newspaper collection is to be extended, ways have to be found to migrate away from the collection of space-hungry, hard-copy newspapers. The newspaper industry already believes that digital is an acceptable long-term storage and preservation medium.

Creating a Sustainable Digital Access Newspaper Strategy

Four potential routes are available for the Library in acquiring an enhanced collection of digital surrogates:

- automatic ingest of PDF format copies of contemporary newspapers direct from publishers, initially by voluntary agreements and then by Legal Deposit legislation;
- systematic digitisation of historic newspapers, in partnership with publishers and content aggregators, both in-copyright and out-of-copyright;
- digitisation-on-demand of other material not accessible via hard copy or microfilm;
- Purchase existing digital licensed collections. The Library already does this, e.g. with The Times historic archive, The Guardian historic archive and The Daily Mirror historic archive.

Automatic ingest of PDF copies

Screenshot of local newspaper digital ingest: