Archiving the ‘daily miracle’

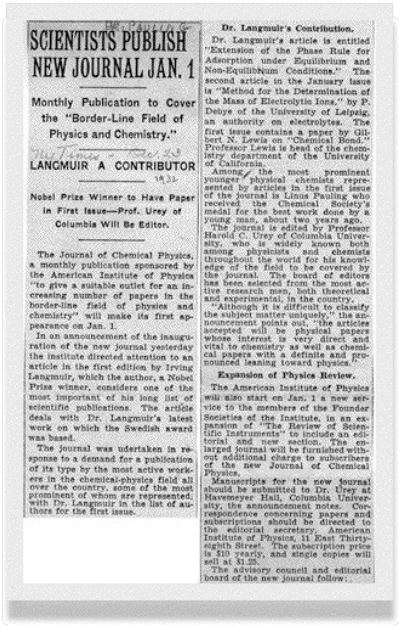
Preservation in the digital newsroom

Blurb

- Ink-stained wretch (Mizzou School of Journalism)
- Librarian/archivist (UCLA)
- 30 years in newspapers
- 25 in newsroom digital technology and access
- Consultant in digital asset management
- Adjunct faculty SJSU-SLIS

The morgue

Where news goes to die.
Useful life:
90 days



SCIENTISTS PUBLISH NEW JOURNAL JAN. 1

Monthly Publication to Cover the "Border-Line Field of Physics and Chemistry."

LANGMUIR A CONTRIBUTOR

Nobel Prize Winner to Have Paper in First Issue—Prof. Urey of Columbia Will Be Editor.

The Journal of Chemical Physics, a monthly publication sponsored by the American Institute of Physics "to give a suitable outlet for an increasing number of papers in the border-line field of physics and chemistry" will make its first appearance on Jan. 1.

In an announcement of the inauguration of the new journal yesterday the institute directed attention to an article in the first edition by Irving Langmuir, which the author, a Nobel Prize winner, considers one of the most important of his long list of scientific publications. The article deals with Dr. Langmuir's latest work on which the Swedish award was based.

The journal was undertaken in response to a demand for a publication of its type by the most active workers in the chemical-physics field all over the country, some of the most prominent of whom are represented with Dr. Langmuir in the list of authors for the first issue.

Dr. Langmuir's Contribution.

Dr. Langmuir's article is entitled "Extension of the Phase Rule for Adsorption under Equilibrium and Non-Equilibrium Conditions." The second article in the January issue is "Method for the Determination of the Mass of Electrolytic Ions," by P. Debye of the University of Leipzig, an authority on electrolytes. The first issue contains a paper by Gilbert N. Lewis on "Chemical Bond." Professor Lewis is head of the chemistry department of the University of California.

Among the most prominent younger physical chemists represented by articles in the first issue of the journal is Linus Pauling who received the Chemical Society's medal for the best work done by a young man, about two years ago.

The journal is edited by Professor Harold C. Urey of Columbia University, who is widely known both among physicists and chemists throughout the world for his knowledge of the field to be covered by the journal. The board of editors has been selected from the most active research men, both theoretical and experimental, in the country.

"Although it is difficult to classify the subject matter uniquely," the announcement points out, "the articles accepted will be physical papers whose interest is very direct and vital to chemistry as well as chemical papers with a definite and pronounced leaning toward physics."

Expansion of Physics Review.

The American Institute of Physics will also start on Jan. 1 a new service to the members of the Founder Societies of the Institute, in an expansion of "The Review of Scientific Instruments" to include an editorial and new section. The enlarged journal will be furnished without additional charge to subscribers of the new Journal of Chemical Physics.

Manuscripts for the new journal should be submitted to Dr. Urey at Mayemeyer Hall, Columbia University, the announcement notes. Correspondence concerning papers and subscriptions should be directed to the editorial secretary, American Institute of Physics, 11 East Thirty-eighth Street. The subscription price is $10 yearly, and single copies will sell at $1.25.

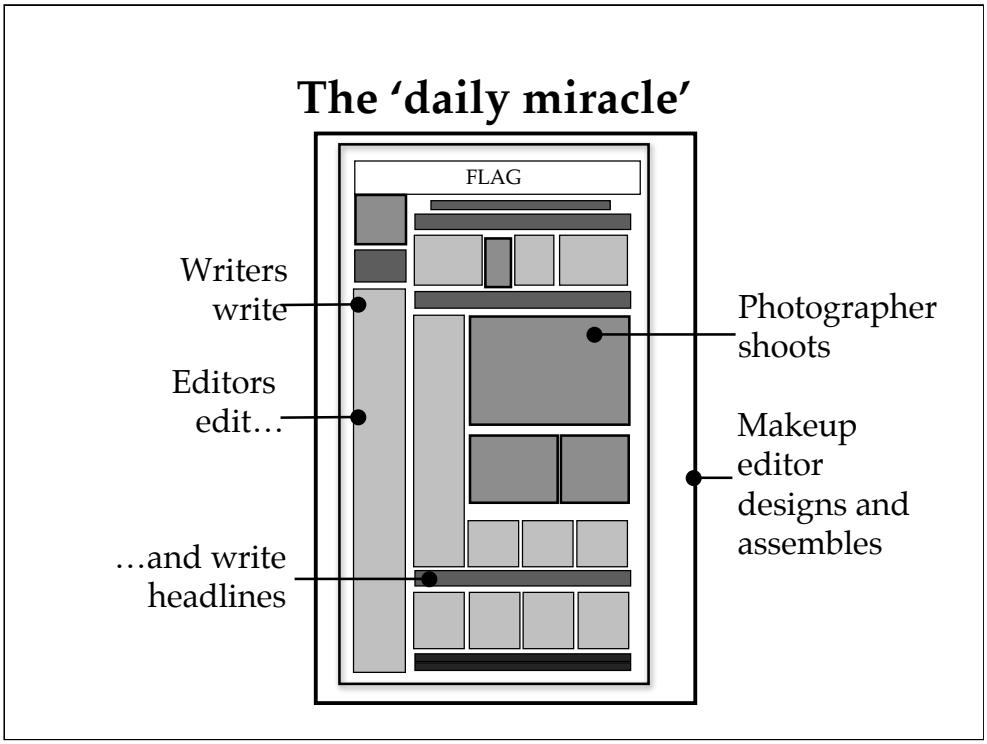
The advisory council and editorial board of the new journal follow:

Former journalist Lee Strobel tells of a young woman from a rural part of Illinois who landed a summer internship on the City Desk of the Chicago Tribune. Her mother, worried about her little girl in the big city, would call frequently to check up on her. One afternoon her worst fears were confirmed when a passing editor happened to pick up the City Desk phone. “Oh, hi, Mrs. So-and-So,” he said. “I’m sorry, you’re daughter isn’t here. She’s down in the morgue.”
The screaming on the other end of the line let that poor editor know a thing or two about newsroom slang. But although the picturesque term *morgue* has faded from the lingo of the high-tech newsroom, news archives are still very much the place where news goes to die.

In fact, Evelyn Waugh had it right in his novel Scoop in 1938, when he wrote that

> News is what a chap who doesn’t care much about anything wants to read. And it’s only news until he reads it. Then it’s dead.

So we have been talking here in New York about something that is very, very ephemeral, not seen as particularly interesting by its creators, and certainly not something worth the massive intervention usually talked about when the subject of digital preservation comes up. News librarians and archivists like to say that interest in a particular story drops off dramatically after a couple of months and is virtually forgotten, old news, dead, after 90 days – until someone needs it. And then they need it *right now* because he’s on deadline.



I hope you have a thorough understanding of why news archives exist: to answer a few basic questions. 1. Have we had this story before? 2. What did we say last time? And 3. What can I recycle from last time? The idea is saving work, not recording history. What's important is speed.

Let's look first at how a page of a newspaper goes together. Writers write, editors edit, photographers shoot. Eventually all of these discrete data objects funnel onto the screen of a page designer, or makeup editor, or paginator – people with big flatscreen monitors and a mouse, who give context and coherence to that day's written record.

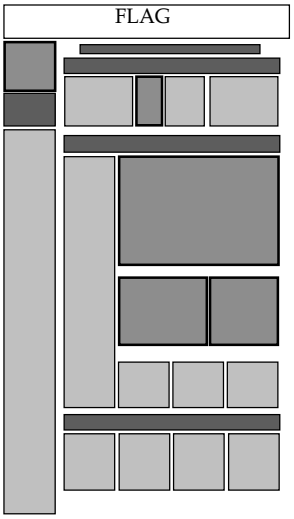
It's a challenging job on unfriendly software. Underneath their page layout tools are big, sophisticated databases performing all kinds of feats like hyphenation and linking and geometry. They have to work quickly and fairly precisely, and they're racing the clock. The accumulated delays of the entire newspaper, from missing ads to balky AP feeds, land on these folks. They are the ones who actually click on the button that represents the decision to publish. Being late costs a lot of money. Every minute a press sits idle involves dollar bills.

Those of you involved in publishing academic journals can play a little thought game: Imagine handling 150 or 200 articles and perhaps a hundred images from editing to print overnight. That's about standard for a big metro daily on Sundays between Thanksgiving and Christmas. That's why putting out a newspaper is called a daily miracle.

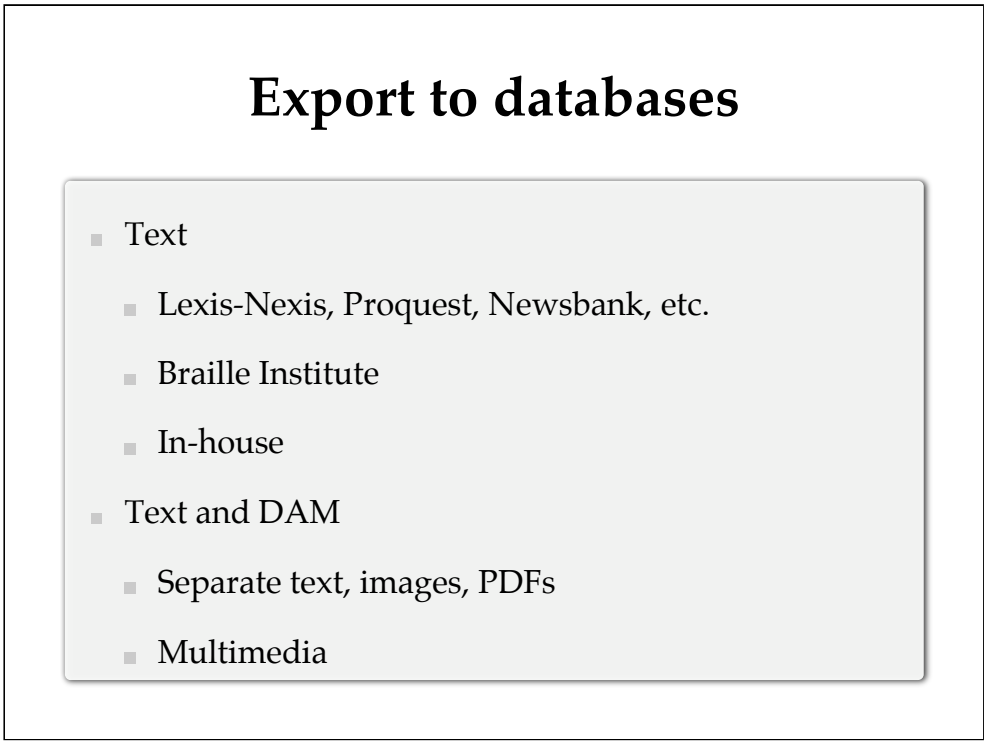
The makeup editor hits the Publish button and lets 'er RIP, as in Raster Image Processor. These big rips turn data into the images of a page, the end result of which is a PDF of the page, a sort of useful waste product of the production cycle.

While the page is ripping, a set of parallel processes is creating that day's archives.

Export to databases



The need for speed … at this stage, news objects are stripped from their matrix, if you will, and parked in systems whose search and retrieval strategies are optimized to the to the type of material at hand at hand.



That is, text archives are extracted for a text databases and will be richly indexed and well structured, and this “submission package” will materialize in a host of places – the big data aggregators, other nonprofit users of text like the Braille Institute and radio reading services, and a newspaper’s in-house archives.

This is a rich package compared to photographs – which will be only minimally indexed but accompanied by rich – and richly misspelled – caption information written by the photographer at the scene.

Export to databases

- Frequently excluded
  - Wire services like Associated Press, Reuters, Agence France-Presse
  - Syndicated material like horoscope, bridge columns
  - Restricted freelance material like op/ed columnists and feature writers

Keep in mind that both formats are now physically separated from the page as a whole. One place you notice this is what *doesn't* show up in the archives or export to the aggregators – all kinds of material like wires, feature columns and stuff whose writers insist that you not keep it.

Where does it end up? Microfilm, if a paper still bothers with it – but that's another speech, another powerpoint…

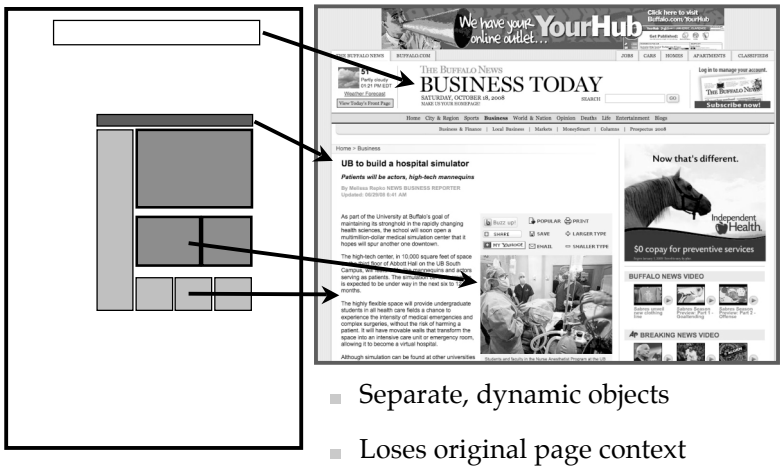
Export to databases

- Bibliographic metadata (pub date, page, author, title, subtitle)
- Authority terms added
- Subsequent corrections and clarifications
- Page context (section front, top story, with pictures)
- Author status (staff, freelance)

To sum up here, the database material ideally has rich bib metadata, use of a controlled vocabularies and authorities, some information that lets you know how a story played, some information about rights, and links to subsequent corrections.

And from here on out, things start falling apart.

Export to databases



- Separate, dynamic objects
- Loses original page context
- Dumbed-down vocabularies

One of the export processes shoots a copy of a story and pictures off to the website. It's stripped of some of its metadata and parked in a generalized category like Business, Sports, Health. It no longer has a connection with the earlier version, and doesn't update dynamically – meaning the archives is faced with supporting or sorting through multiple versions.

Archiving web pages is a daily struggle for many newspapers, and they simply don't do it.

Export to databases

- Imperfect process
  - Bib data missing or erroneous
  - Corrections not linked
  - Rights information unavailable or unknown
  - Objects missing from PDFs
  - Separate Web version


Kuwait Braces for Iraq's Wrath

Some citizens not as concerned

An Uneasy Quiet on Lebanese Border

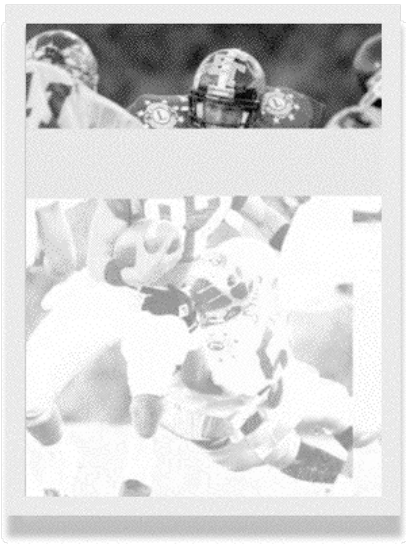
I said earlier that makeup editors must work quickly and precisely. But think about it. If it is necessary for the integrity of the archives and digital preservation that you use a rigid sequence of steps and never deviate from procedure, you'll never put out a paper. These systems are designed to be flexible, to allow workarounds. You have to meet your press deadline.

Unfortunately, workarounds play a certain amount of havoc with databases. This is a problem that Newsday encountered several years ago that resulted in stories dropped from the PDF. They attributed it to a workaround, but could never replicate the problem reliably.

The economic tradeoff between clean data and cost used to tilt toward clean data, but those days are over, as we'll see. So when you look at a text or PDF database, know that there are some serious deficiencies, a few of which I've listed here.

L.A. Times photo archives, 1997

Stored, backed up
Minimal terms applied
Files counted but not validated
Migrations not documented
About 1% corrupted (3,000 images)



Once material is in a database, the tried-and-true principles of benign neglect take over. Typically, if you ask newspaper IT folks about preservation, they'll assure you they're backing it up – although you'd be surprised at the number of newspapers that don't back up their data.

Storage backup is not preservation. A couple years ago we ran some tests on about 300,000 10-year-old images in the image archives database at the Los Angeles Times. It wasn't real scientific, just a statistically valid sample of pictures that we ran through three filters that analyzed some of the information in the JPEG header.

What we saw in just under 1% of the files was something like this, missing data, color shift, images chunked up and moved.

L.A. Times photo archives, 1997

Stored, backed up
Minimal terms applied
Files counted but not validated
Migrations not documented
About 1% corrupted (3,000 images)

This is a 100-year-old orange tree, photo taken in 1997.

L.A. Times photo archives, 1997

Stored, backed up
Minimal terms applied
Files counted but not validated
Migrations not documented
About 1% corrupted (3,000 images)

My favorite, three girls in a college Shakespeare production. Let's zoom in…

Three things going on here. Picture is sliced, diced and moved around, like a puzzle. Because backups and media refreshment or software upgrades aren't well documented, there's no way of knowing when, how or why these problems popped up.

We did find out that these errors are associated with a header error that says "some number of extraneous bytes at marker zero X D 9." This is apparently a bytestream within the JPEG software that tells the picture to start rendering.

Of course, if you aren't looking routinely for this sort of error your periodic backups will overwrite bad images over the good ones. This is exactly what happened in the Times databases.

One percent is only a few thousand photos, and probably a tolerable loss rate. But newspapers aren't defining "tolerable" or routinely looking out for it.

L.A. Times graphics archives, 1995

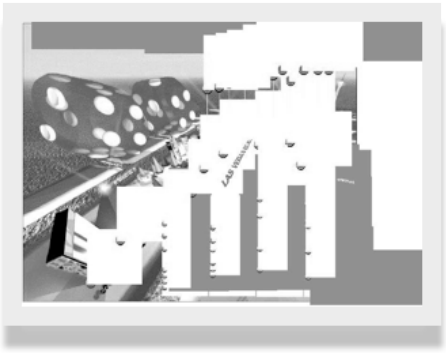
Stored, backed up
Minimal metadata
Obsolete software
Keyword searchable, retrievable
Files can't be opened

Here's another kind of problem, technology obsolescence.

These are vector files that are several software versions and a Motorola microchip away from being viable. There are tens of thousands of them in the Times database. This graphic described the effects of brushfire and torrential rain on L.A.'s sliding hillsides. Interesting that the colored areas you see are raster images embedded in vector software. Vector software is highly problematic from a technological standpoint. What I find interesting about this problem is that there is no systematic way of deleting them from the database so they'll occupy storage space forever, or until someone decides to pull the plug on older files without really knowing what they are.

L.A. Times graphics archives, 1995

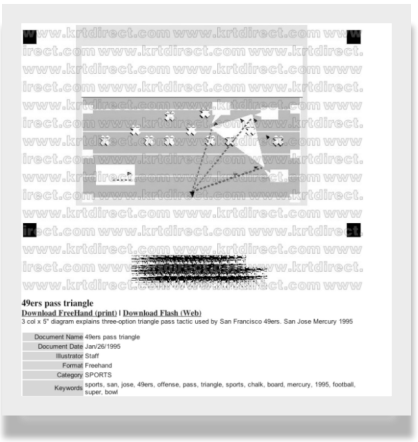
Stored, backed up
Minimal metadata
Obsolete software
Keyword searchable, retrievable
Files can't be opened



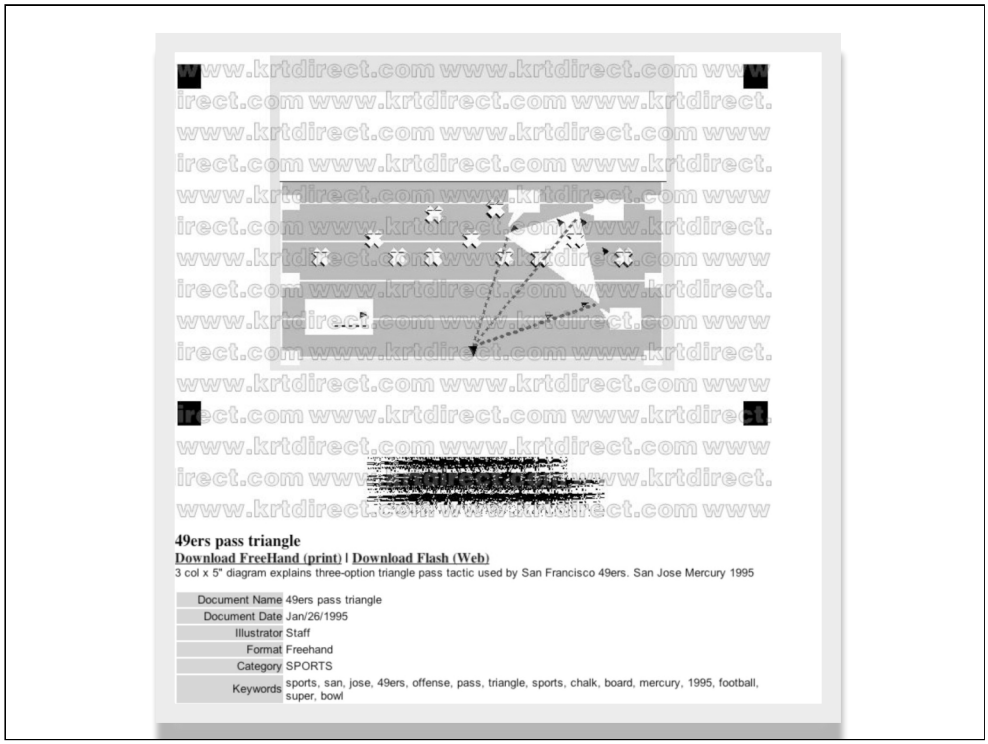
Here's another graphic describing the building boom in the late 1990s on the Las Vegas Strip. We seemed to have preserved the dice, anyway, but the text is absent.

L.A. Times graphics archives, 1995

Stored, backed up
Minimal metadata
Obsolete software
Keyword searchable, retrievable
Files can't be opened



The Knight-Ridder-Tribune graphics wire service is even selling this stuff. This is a 1995 graphic that as far as I know still lives in its database.

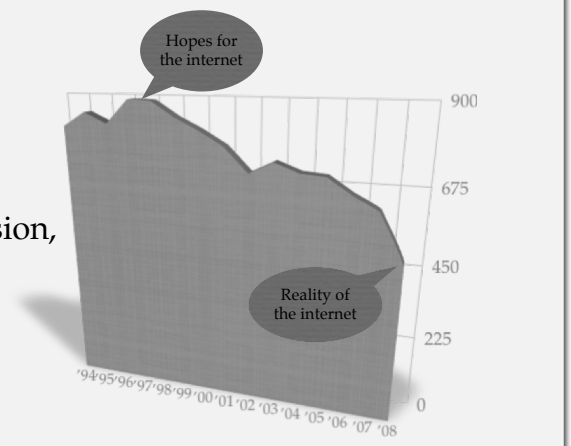


The image survives – I think this is a 1995 football graphic – because you're looking at a PDF. The horizontal black blob is the text, which collapsed because the imaging software couldn't properly render the font.

I don't know that KRT is systematically weeding its database. News organizations typically don't bother, accumulating vast terabytes of data through what one of Associated Press's technologists calls "organic convenience."

The archivists

- Membership in SLA News Division, 1994 - 2008



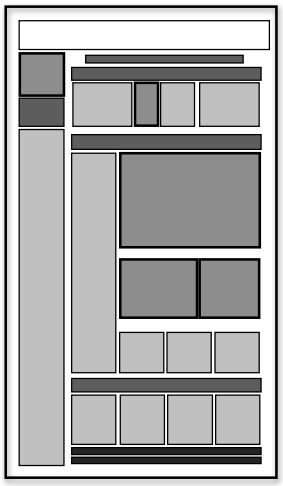
As John Carroll told us, the internet has created some tremendous opportunities and serious threats to print newspapers. One thing no one saw coming was the rapidity with which print advertising collapsed and migrated elsewhere. When management looked around for places to cut costs, the news libraries and archives were, in the jargon, low-hanging fruit.

So now, as we look to tackle digital preservation, the people who would have been there to talk with us about data standards, best practices, format migration and the like – are gone, and the ones who remain are working overtime trying to publish tomorrow's news.

I bring up newspaper economics and its effect on the archives because those people were and are the crucial layer between daily chaos and well-formed data structures in databases. Do not think for a minute that Nexis, Newsbank and Proquest are vetting all the data at ingest. Next time you see a random word or nonsensical character string where you expected a headline, it's because there was no one to clean up after the automated processes. The net result is an accumulation of dirty data both in the aggregator's data stores and in the newspapers' own archives.

News archives today

Aging databases
Dwindling quality control
Orphan metadata
Orphan works
'Storage' mindset
'Waxy buildup' in bottomless databases



Evelyn Waugh wasn't very nice on the subject of the morgue, referring to the "detailed inventions "and "intricate misrepresentations" of daily reportage. I'm assuming that all of us here care about keeping this stuff. What are the threats?

•Aging databases and the gradual bit here, byte there corruption that happens even if you are paying attention.
•Fewer people and more algorithms devoted to catching errors. You need very inventive algorithms to keep up with page designers and makeup editors.
•Orphans, orphans everywhere – pictures and articles by freelancers that don't get kept, venerable vocabularies abandoned in the dumbed-down categorizations of the web, contextual metadata with no place to go.
•A storage mindset that overlooks what' s required for long-term access by human users.
• Bottomless databases with obsolete or inaccessible information measured in terabytes. Do photographers *really* need to keep every image they shoot with their digital cameras?

-30-

Vicky McCargar

mccargar @ mac.com

Will print newspapers go away? Will web publishing take over? Can the born-digital processes be successfully automated? Is microfilm dead? Is digitizing microfilm setting up another set of hard-to-preserve digital objects (the TIFFs)? These are all huge preservation questions.

There are some organizations that look a bit like repositories – Associated Press and Lexis Nexis, for example. They'll be the first to tell you that's not their business, but there are some successes there that are worth exploring. Newspapers need help, but I'm not sure they even know it.