From Stockwell to the web (via Tel Aviv)

Today the Guardian launches a UK national newspaper first – a searchable digital archive which will soon contain all the copies of the paper and its sister title, the Observer, from their first issues in 1821 and 1791. To celebrate this momentous step we are publishing a series of special supplements. Here, Oliver Burkeman explains how acres of dusty, yellowing pages were translated on to the internet

Nobody really knows how much information there is in the world. According to one extremely rough estimate, if you took every book, newspaper, magazine, TV and radio programme, every music album, every handwritten letter, every filed-away document and every other piece of recorded data in existence, and you stored them all on computer hard drives, the amount of disk space you would need would be somewhere in the region of 2,100 exabytes, or 2,100bn gigabytes. If it helps – and it probably doesn't – this is more than 100m times the amount of data that is thought to be stored, in print form, in the bookstacks of the world's largest library, the Library of Congress in Washington.

What is certain is that only a tiny proportion of all this information is currently available on the internet. We joke darkly that Google knows everything – but even Eric Schmidt, Google's chief executive, admits that it probably knows less than 10% of everything that is publicly available to be known.

Putting all the world's books and old newspapers and broadcasts on the web is a project of almost unimaginable scale: Schmidt has estimated that it might take another 300 years. Still, a few weeks ago, on the 12th floor of an office building just outside Tel Aviv, a small group of Israeli technologists finished making a modest contribution to that task, by digitising every page of every edition of the Guardian since its launch in 1821 up to the 1970s, along with thousands of editions of the Observer – about 900,000 pages in all.

Yoni Stern, co-founder of Olive Software, which carried out the work, is a man of diffident bearing, but his ambitions are not exactly restrained. "For years, I kept thinking, well, what about all this content that isn't on the internet?" he said on a blazingly hot day in Israel last week. "Maybe, one day, all of the content in the world that was ever created will be connected online. And maybe my destiny is to help this process of connection."

The Guardian Digital Archive, launched today, allows users to search the full text of the newspaper from 1821 to 1975, and the Observer from 1900 to 1975 – the first time a national UK newspaper has made its paper archive available to the public online. The second phase, early next year, will see the addition of the Observer from 1791 to 1899, and both papers from 1976 to 2003. Instead of merely calling up the text of news stories, the system enables users to view articles as they appeared in the newspaper, then to browse the rest of that day's edition. And so you can, should you wish, discover, on the same front page of the Guardian that reported the abolition of slavery, an advertisement for Lessey's Carbonated Seidlitz Powder ("recommended for the relief of the nausea, heartburn, debility of the stomach, and other symptoms induced by the heat of the weather").

One side effect of the internet is that it can be almost too easy to find exactly what you want, and to ignore everything else, eliminating the pleasure of finding things you didn't know you were looking for. The digital archive permits precise searching, of course, but being able to browse entire editions is addictive; it makes searching a newspaper archive far more similar to the experience of reading a newspaper. Recently, I've been following a dastardly tale of postal-order forgery in 1960s Manchester – a classic cat-and-mouse game between the police and a shadowy criminal mastermind – but only because I found a report on it a few pages after Alistair Cooke's front-page splash about the assassination of JFK. "The real brains behind this is a man called John McAllister. I have not yet been able to trace him," a frustrated Detective Inspector C Mellor grumbled to the Guardian at the time.

There is nothing new about computer software which can recognise words on the printed page. This is the technology behind Google's controversial effort to scan and make searchable at least 15m books at a reported rate of 3,000 a day, including much of the stock of the Oxford and Harvard university libraries. But books are simple things, full of uniformly designed slabs of continuous text. Newspapers, by contrast, are a cacophony of columns, headlines, adverts, photographs, captions, crosswords and weather maps; in earlier centuries, there wasn't uniformity in the width of a column, or the size of a headline, from one page to the next. If you can't tell all these elements apart, it's impossible to build a useful archive.

Olive Software's secret ingredient is its system of "componentisation" – a set of mathematical algorithms that allow its computers to learn how to make sense of the cacophony and thus, in effect, to learn how to read a newspaper. The system acts as a bridge between the electronic realm and the realm of printed words on paper, which Stern has an alarming tendency of referring to as the "legacy world". Most of the other firms considered for the job by the Guardian proposed a different method of componentisation: they planned to use hundreds of human beings to plough, bleary-eyed, through digital images of every page of the paper, marking by hand where one story stopped and the next one began.

Few places belong more definitively to the legacy world than the Stockwell Deep Level Shelter, a former air-raid shelter so far beneath the streets of south London that you can hear underground trains rumbling past overhead. At ground level, it doesn't look like much: a circular white concrete pillbox building with a blank metal door. But behind the door is a lift-shaft, and at the bottom of the lift-shaft – reached using an old-fashioned cage lift, highly unsuitable for claustrophobics – is an extensive network of tunnels that once accommodated up to 8,000 people at a time.

The shelter opened in 1944, offering protection from V-1 flying bombs, and was later used as a camp for soldiers en route through London. It isn't open to the public, and stepping into the two main tunnels, each nearly half a kilometre long, is unsettling: you get the irrational but distinct feeling that the voices of its wartime occupants might only just have died away. Today, the shelter is operated by the archiving company Recall Total Information Management, which houses many of its customers' boxes of documents on the original six-person metal bedframes. Among the collections are around 250 leatherbound volumes of the Guardian and Observer, dating back to the early 1800s. (Other volumes are stored in the Newsroom, the paper's archive and visitor centre.)

These very oldest copies demonstrate that paper doesn't last forever. Even in Stockwell's dark, cool conditions, the archives are slowly disintegrating: gingerly opening one volume from 1846, I released a cloud of acidic dust which settled all over my clothes. (Underneath the dust, from the top of the front page of Saturday, July 18, 1846, where these days the lead story would be, it read: "Found, a purse, which the owner may have by describing it. Apply at the Clarence Hotel.")

"If you put a copy of today's Guardian in the strong sunlight for three days, it'll no longer be white," says Ed King, head of the British Library's newspaper collections at Colindale in north London. "What is happening in libraries and archives, as a general phenomenon, is that paper is degrading – more slowly, but the process is normal and inevitable. Controlled temperature and humidity can slow it down, but not eliminate it entirely."

Browsing the archive can be addictive. I've been following a dastardly tale of postal order forgery in 1960s Manchester

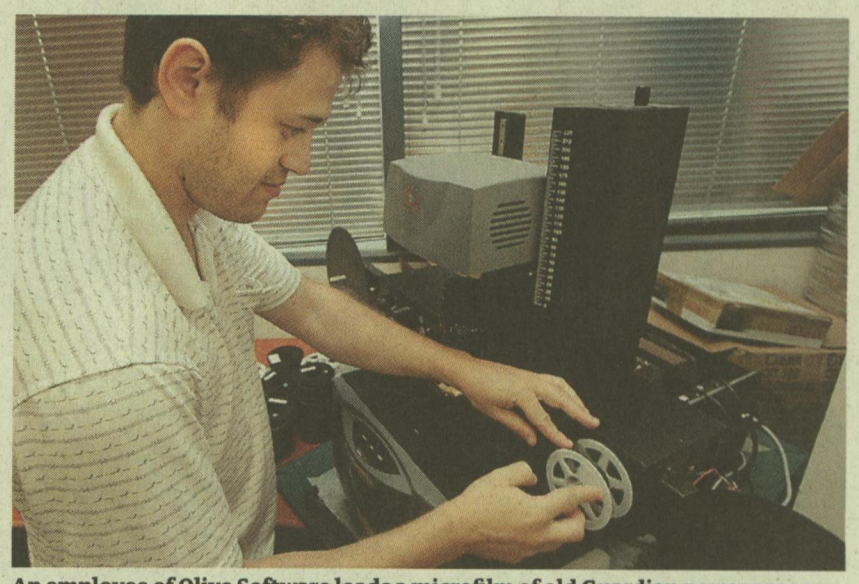
Counterintuitively, the oldest copies of the Guardian are in a better state than some newer ones, because older paper was made largely of fibre from rags. But overall the situation is grave. Fifteen per cent of Colindale's historic newspapers already can't be touched by the public, because their condition is too poor.

From the 1950s, libraries and newspaper firms thought they had found the solution: microfilm. In his 2001 book Double Fold, the novelist Nicholson Baker launches an unlikely tirade against the practice, accusing an overzealous band of library modernisers, mainly in the US and Britain, of having sold off or destroyed countless shelves of books and newspapers in pursuit of their crazed vision of libraries consisting of rack upon rack of neatly stored rolls of cellulose acetate.

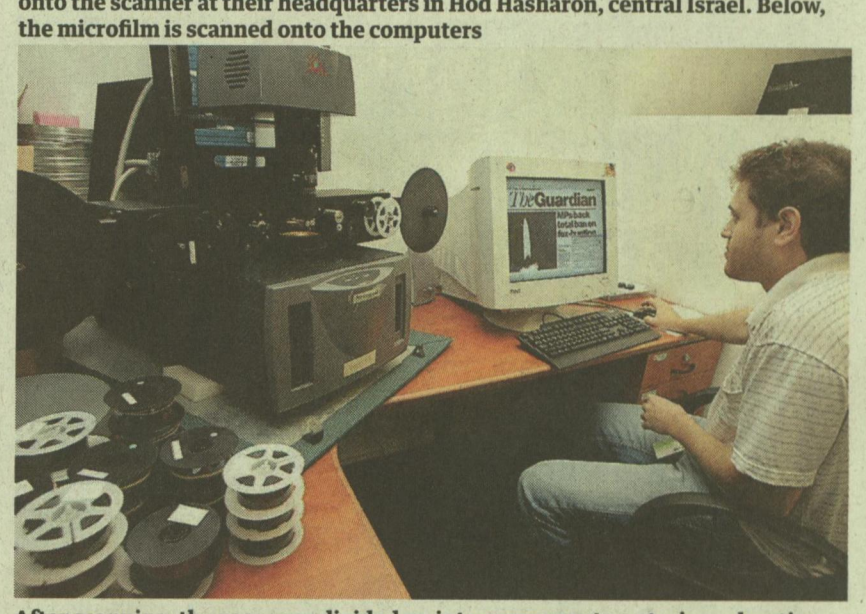
The "problem" of libraries, said the leading American microfilm crusader of the 1950s, Verner Clapp, was to prevent the past from "clogging the channels of the present". His ideas led to the establishment of America's $358m (£172m) Brittle Books Program, which saw thousands of books ripped from their bindings in order to be photographed.

Baker's book makes plain the limitations of microfilm: it is fiddly to use, the film stock itself is fatally subject to the effects of ageing, and reading a newspaper on microfilm is

The underground storage facility below Stockwell tube station, south London, where old copies of the Guardian are kept
Photograph: Felix Clay



An employee of Olive Software loads a microfilm of old Guardian newspapers onto the scanner at their headquarters in Hod Hasharon, central Israel. Below, the microfilm is scanned onto the computers



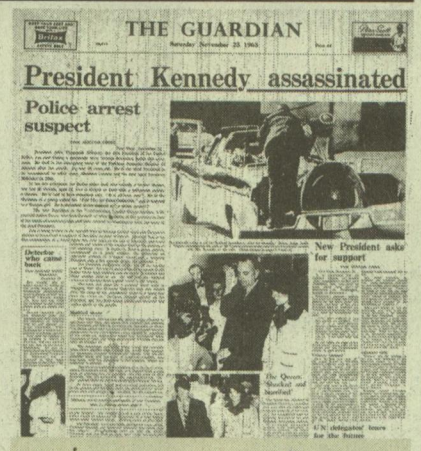
After scanning, the pages are divided up into components – stories, advertisements, headlines and pictures. This is an 80% automatic process, overseen by an operator to ensure accuracy Photographs: Gali Tibbon

nothing like reading one on paper. But even today, it is the only standardised, agreed-upon remedy for dealing with the problem of disintegrating paper, which is why the British Library, despite having launched a major web archive of its 19th-century newspapers, continues to use microfilm. It would be too risky to stop, Ed King argues, while we still can't be sure that any given digital format is really going to last. Try accessing the data on a 5.25-inch floppy disk these days, or from a file created in the defunct program WordPerfect. And so the newspapers arriving at Colindale each day are dutifully photographed – after having been ironed, to get rid of the folds and crinkles.

The Guardian still uses microfilm, too. A full run of the paper is maintained and updated by ProQuest, a Michigan-based company, and it was copies of those rolls of film that were shipped to Israel late last year, to begin the mammoth task of putting the newspapers' archives online.

One of the hazards of working for Olive Software, which is busily digitising the archives of scores of newspapers from the US, South America and Europe, is the risk of getting distracted. You might sit down at a computer to adjust the image sharpness on a batch of digitised pages, only to find that it is two hours later, and that you've spent the intervening time immersed in coverage of the Battle of the Somme. "My first day of training, I got through two to three pages in the whole day," said Chezkie Kasnett, a sales manager. "I was, like, 'This is just amazing stuff!' I was reading the Irish Times, I think, right around 1942 – the invasion of the allies. Detailed descriptions of the battles. You can very easily get into the material." Partly as a result, the operators who oversee the automated process on a day-to-day basis, in two shifts of eight hours each, are trained to understand the structure of a newspaper without reading the content – to see it, like the computer does, as a smorgasbord of jostling components. "If you had to start reading material to understand it, a project would take us decades," said Kasnett. "We did have one operator who would read every newspaper," his colleague, Yael Arbel, recalled. "He is no longer an operator."

Olive likes to describe its process as a hybrid of human and machine – 80% automated, but with real people providing a level of accuracy that computers still can't attain. Their offices, in a palm-lined business park in Hod Hasharon, north-east of Tel Aviv, are essentially a conveyor-belt for information. The rolls of microfilm are scanned onto computers by human operators, then cleaned up by computer with human assistance. Then – the heart of the process – the pages are divided up into components in Olive's password-protected server room, where banks of processors hum 24 hours a day, processing up to 1m pages a month. What emerges at the other end are images of newspaper pages covered with translucent rectangles of different colours, each intended to identify a different story, advertisement or other element. These images are inspected by still more humans, to correct the computers' mistakes.


THE GUARDIAN
Saturday November 23 1963
President Kennedy assassinated
Police arrest suspect
New President asks for support

How to access the digitalarchive

The archive is available as an online subscription service at www.guardian.co.uk/archive. Readers can purchase timed access with unlimited downloads. There are three options: 24 hours (£7.95), 3 days (£14.95) and 1 month (£49.95). Longer periods can be purchased on request by emailing archive.help@guardian.co.uk. During launch we are offering free 24-hour passes to all Guardian readers. To redeem your free pass, please visit the website. If you want to come back for more, there is also a 50% discount on all passes. This introductory offer will end on 30 November 2007. If you want to find out how to get access for your school, university or library, please contact syndication@guardian.co.uk.

Text recognition comes last. For 20th and 21st century newspapers, the process is relatively simple: when I visited, the system had no problem reading almost every word of a Guardian front page featuring reports from 1950s Rhodesia by the foreign correspondents Hella Pick and Ian Aitken. For older editions, it is more of a challenge, and for some European newspapers, printed in Gothic text, the character recognition software has had to be significantly rewritten.

The finished data is stored in standard graphics files and in the simple, general-purpose, open-access computer language, XML – part of an effort to increase its lifespan as technology changes. The sheer volume of data required that the Guardian archive be shipped to the paper's headquarters in London on hard drives, not DVDs.

Yoni Stern's co-founder is a professorial Russian computer scientist named Emil Shteinvil, who emigrated after the collapse of communism with three children, no ability to speak English or Hebrew and, he says, $300 in his pocket. Like the firm's other senior staff, he makes no pretence of being interested solely in the technical side of things. "This is history, OK?" he said, over lunch in the building's communal canteen. "You can feel all this life, and this is very interesting! You can see all about this war, this election, this day! You can see a piece of writing on how to care for your slaves!"

For Stephen Marks, the London-born executive who runs Olive's Israel offices, the digitisation of the world's newspaper archives promises nothing less – though he is biased, naturally – than a transformation of the concept of history. "History is written by the victors, and then history books are authored, and then they're put on a syllabus," he said. Conventional wisdom gets entrenched. But give people direct access to a multiplicity of viewpoints from the time of the events themselves, and everything changes. "It's not that newspapers don't have an ideology; it's that there are lots of them. The bombing of Dresden: how was that being reported from the UK? From Dresden?"

"A newspaper today might be full of bullshit," says Yoni Stern, "but it's all a part of the history of culture – the bullshit, too, no less than the reality. You can't get that from history books."

From Olive's offices, you can look down over eastern Israel, across to Palestinian villages and Jewish settlements in the West Bank, and a chainlink portion of the hugely controversial security barrier. When you do, it suddenly seems no coincidence that this technology reached its zenith in Israel, where conflicting versions of history permeate present-day life so completely. Digitisation represents "a big opening-up of history," Marks said. "For the last 350 years or so, we've recorded our history by the day, from all sorts of different viewpoints, but it's always been hard to get at before now. This is like the Dead Sea Scrolls for the last few centuries of humanity."

The Guardian, so far, has documented 186 of those years. We invite you to start exploring.