CHAPTER 12

ARCHIVAL TRANSITIONS: SOME DIGITAL PROPOSITIONS

PELLE SNICKARS

IN MID-JULY 2009 Wikinews published an article claiming that the National Portrait Gallery in London threatened a U.S. citizen with legal action since he had allegedly breached the museum's copyright of several thousands of photographs of works of art. Apparently, the young American Derrick Coetzee had come up with a program that automatically would download high resolution imagery from the National Portrait Gallery's website, images that Coetzee as a regular contributor to Wikimedia Commons uploaded to the site. Wikimedia Commons is the database of free-to-use media that Wikipedia writers use for illustrations; today the database contains some five million photographs. The digital images that Coetzee uploaded were exact digitized reproductions of artworks, drawings and older photographs. Since the holdings of the National Portrait Gallery consisted of mostly older material, Coetzee considered the digital reproductions to belong to the public domain and thus free for public use under United States law (where he and Wikimedia Commons were based). The crux of the matter was that copyright to digital reproductions was claimed to exist in the U.K. where the museum was situated. Hence, in "a letter from [the museum's] solicitors sent to Coetzee via electronic mail, the National Portrait Gallery asserted that it holds copyright in the photographs under U.K. law."[1] They demanded

that Coetzee provided undertakings to remove all of the images from Wikimedia Commons.

Derrick Coetzee uploaded the e-mail to his Wikimedia Commons account, and since Wikipedia has 330 millions of users news of the event spread quickly. BBC picked up the story—and soon the National Portrait Gallery had most of the blogosphere and numerous web commentators as its opponents. In Britain, copyright law apparently gives new copy-

FIGURE 1: Elizabeth I of England, the Armada Portrait 1588. Digital image gathered by Derrick Coetzee from the National Portrait Gallery and uploaded to Wikimedia Commons which states: "While Commons policy accepts the use of this media, one or more third parties have made copyright claims against Wikimedia Commons in relation to the work from which this is sourced or a purely mechanical reproduction thereof."

right to someone who produces an image full of public domain material, Cory Doctorow at boingboing.net sarcastically commented, "effectively creating perpetual copyright for a museum that owns the original image, since they can decide who gets to copy it and then set terms on those copies that prevent them being treated as public domain [...] If you take public money to buy art, you should make that art available to the public using the best, most efficient means possible."[2] Under the heading "Who's Art Is It, Anyway?" *The Wall Street Journal* wrote that it was not hard to understand the museum's frustration. "It goes to all the trouble and expense of making accurate photographic copies [...] and then someone comes along with a few clicks of a mouse and appropriates thousands of images." However, copyright law tries to balance between private ownership of intellectual property and public usage decades after the work's author is dead. "If new copyrights can be attached to old works of art, the whole copyright system is thrown out of whack." According to the *The Wall Street Journal* copyright law exists, and has existed for one purpose: to make creativity pay. Producing exact photographic copies of paintings "is no doubt valuable and involves painstaking work. But it isn't—and isn't meant to be—creative." The champions of intellectual property, the newspaper asserted, "can't afford to waste their energies trying to monopolize images that already properly belong to us all."[3]

HERITAGE INSTITUTIONS AND THE WEB

During the last decade heritage institutions across the world have been challenged by new digital technologies as well as by entrepreneurs who rapidly understood the advantages of the networks of networks characterizing the Web. The discussion and media debates following the polemics between Derrick Coetzee and the National Portrait Gallery is, hence, not only a clash between a new binary American frontier and an old European museum. It also highlights how memory institutions within the so-called ALM-sector (archives, libraries, museums) have been seriously contested by unprecedented cultural players that embraced new information technology, given them as much popularity as conceptual advantage. Heritage institutions have for long been remarkably absent, and at times almost invisible on the Web. The situation is beginning to change, but during the late 1990s the absence resulted in a situation where the digital

domain was largely left to the market, and various user-oriented initiatives as Wikimedia Commons. In 1995, for example, when the two major American online photographic archives were established on the Web, Corbis and Getty Images—taking advantage of the possibility of using binary code as the new interface for older stock imagery—most European heritage institutions were only beginning to think about whether or not one should upload metadata on the Web. Naturally, budgets were tighter and copyright put numerous restrictions on material, yet the conceptual understanding of digital technologies was by large absent. Of course, there were exceptions. A number of American heritage institutions for example, notably the Library of Congress, were keen on using new digital technology. Yet as the controversy above illustrates heritage institutions across Europe are still, fifteen years later struggling with how to perceive and conceptiualize the Web.

In this article I would like to address the relation between the contemporary Web and heritage institutions in a broad sense. Various American websites and companies have during the last decade in many ways challenged the ALM-sector, apparent not the least from a European perspective. In short, doing American studies within the memory sector simply means acknowledging that U.S enterprises like YouTube, Flickr or Wikipedia have paved the way for completely new ways of organizing heritage material for the global user. The purpose of the article is, hence, to outline a number of archival transitions, as well as put forward and discuss some digital propositions focused on how heritage institutions could, or rather should understand and use the Internet as well as the Web. The "digital" is not a threat to archives and the ALM-sector—it is a blessing. Digital user patterns are totally different from analogue ones, social tagging can produce the most amazing results, and through various forms of data mining digitized collections of heritage material can be used to deduce unprecedented knowledge structures. The networkable nature of new digital technology can also be applied to various distributed digital preservation programs—in short, file sharing as a storage model in the age of the networked archive.

In fact, the most challenging task facing heritage institutions today is how to deal with the new binary networks of networks. On the one hand, the Web needs to be thought of as both a tool for access, primarily for distributing collected heritage material, as well as distributed preservation. But on the other hand, the Web also poses numerous problems for

heritage institutions, not the least in terms of collection and assemblage strategies of new digital material. Access to information is what the Web does best, and although controversial, the Google Book Search project vividly illustrates how heritage material could be accessible on a mass scale. Today more than two million books belonging to the public domain are downloadable from Google Book Search in full PDF-format, and numerous national book services providing information on titles held by university and research libraries, as for example the Swedish Libris system, uses the open APIs from Google to include links to these scanned books. An API [application programming interface] is an interface that defines ways by which an application program may request services or data from another system. Open APIs have, in short, changed the way in which private web-based companies as well as publicly funded institutions interact on the Internet by exchanging information and data using the network as a distribution channel for Web services.

GOOGLING FOR BOOKS

What the Google Book Search project has foremost taught heritage institutions is that the costs of digitizing per item can be radically cut if one digitizes in an industrial manner rather than in a traditional way based on choosing material on certain criteria. In the Google Book Search nothing is choosen; instead stack after stack is digitized, and like everything else at Google the project is based on scalability. Mass digitization, then, concerns millions of books rather than millions of pages. Even though many heritage institutions "welcome the unprecedented access to all this information, Google has also been criticized for the inferior quality of their images, the emphasis on the English language, the violation of copyright laws and the lack of attention for preservation issues", to quote the librarian Astrid Verhausen. "The question therefore arises", she writes, "can libraries do better than Google?"[4] Well, of course they can in terms of quality, and they did during the 1990s—but never ever in terms of quantity. The reason is scale, money and know-how. One of the partners in the Google Book Search project, the Bavarian State Library in Munich, will receive approximately 50 million Euros from Google, a sum most national libraries can only dream off.[5] "The digitisation of books is a Herculean task but also opens up cultural content to millions

of citizens in Europe and beyond," Viviane Reding, EU Commissioner for Information Society and Media, recently stated in a press release. However, regarding Google and the U.S. lead and advantage in the digitization struggle, he also confessed that "that Member States must stop envying progress made in other continents and finally do their own homework. It also shows that Europeana alone will not suffice to put Europe on the digital map of the world. We need to work better together to make Europe's copyright framework fit for the digital age."[6]

Still, the Web with its abundance of information also poses problems for heritage institutions. The legal deposit law for example doesn't work at all in the digital domain, and Web content characteristics as high dynamics and format variety makes Web archiving a real challenge. Since the Web 2.0 revolution most information seems to have dispersed into bits and bytes that are constantly changing. Preserving content that is bottom-up driven, distinguished by interactivity and a networkable nature is problematic — if not impossible. Heritage institutions should acknowledge this, not the least in a situation where the cloud computing trend is transforming parts of the Web into a platform of distributed storage. The new cultural landscape of the 21th century is, hence, very different from the previous one. It is made of binary code, and we are increasingly living in a software culture, "a culture where the production, distribution, and reception of most content — and increasingly, experiences — is mediated by software."[7]

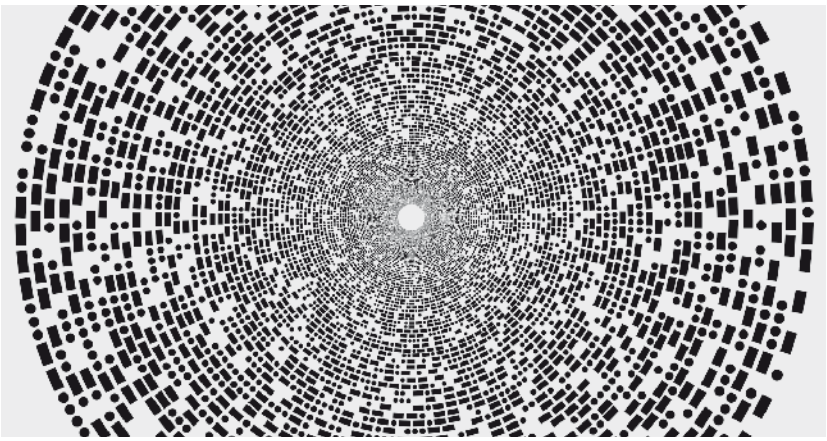

FIGURE 2: All cultural systems of modern society run on software—binary code is the invisible glue that ties everything together.

The software culture regulating the Web is ubiquitous — it is everywhere. As the frequency of updates and mash-ups increases, it is for example more and more difficult from an archival heritage perspective to pinpoint what a "digital object" actually is — even if it is conceptually perceived as the sum of its parts. When the digital object is interactive, it constantly changes, or as media theorist Lev Manovich has stated: "When a user interacts with a software application that presents cultural content [as for example Google Earth] this content often does not have definite finite boundaries. [...] Google Earth is an example of a new interactive 'document' which does not have its content all predefined. Its content changes and grows over time."[8]

Yet, at a time when "the digital" has become the default value of culture, most heritage institutions are still hesitant in responding to the new situation. As is well known, we are at the moment undergoing a shift in cultural production, consumption and distribution of more or less cataclysmic proportions. The lines can be drawn between what Lawrence Lessig has called the "Read-Only" and the "Read/Write cultures"—RO and RW respectively.[9] From a heritage perspective, previous analogue RO media had sort of an "object" quality. In the film archive, for example, archivist could store film reels in the vault, and even though a film was often made up of several reels, they were still objects to be preserved on shelves. The matter totally changed with the introduction of binary code and the coming of digital and immaterial media. Within the contemporary Read/Write cultures, documents have begun to loose their static boundaries. Naturally, digital media for example can still be an object—a CD or DVD for example — but online RO media and the new cultural forms visible on the Web gradually resemble kind of processes. If "production" was a code word for the 20th century cultural landscape, "distribution" seems now to be the buzzword of the new software culture.

ARCHIVAL MODE OF ONLINE MEDIA

Being online, media scholar Geert Lovink has claimed, "we no longer watch films and television—we watch databases."[10] The digital archive is by nature a database, that is, a structured collection of data stored in a computer system. Database structures are organized according to various models: there are relational ones, hierarchical ones, networked ones and so on

and so forth. Focusing databases, however as central for new media, is hardly a novelty; it is after all the main theme of Lev Manovich's book *The Language of New Media* written almost ten years ago.[11] Nevertheless, a consequence of the database structure of online media, which arguably wasn't too visible a decade ago, is that the differences between various media forms seem to be disappearing. Inside the media archive (or within the database), the concept of medium specificity is starting to become archaic. On the Web all media seems to be gray; or more correctly, on the Web there are on closer observation no media at all, just files in databases containing numerically coded information. Just as the 20th century media forms are converging they are replaced by surface effects of algorithms, that is, by various kinds of programmed content consisting of text, sounds and (moving) images. Filled as it is with binary files, the Internet would seem to be the only channel of communication that still remains. Access to media history and memory, hence, becomes a question where to look.

At the same time media history teaches us that new media never radically replaces older media forms. As a consequence a rather strict division between different media forms still prevails on the Web. For instance, when public service radio or television has been upgraded to digital platforms, the programs are still packaged using the respective media's special signatures, logotype etcetera. Web based television is still seen as an *extension* of conventional TV—even though this may gradually be changing within the industry. The specificity of the medium is still rooted in the analogue past and not in the digital future. Naturally, there are exceptions; pod casting for example has become a distinct media-specific feature of online radio. Different from conventional radio, but similar to newspapers and magazines, pod casting allows listeners to subscribe to programs. They can be automatically downloaded to one's mobile Internet device, and like a book or newspaper used whenever.

However, a major difference between analogue media and its subsequent binary online version is the latter's database structure—or its "archival mode." File clusters shared at various P2P-networks are a good example of the archival mode of online media. The Pirate Bay in Sweden (or wherever it is now located) has for years had more than one million trackers to various media files and more than 25 million unique peers. Torrents themselves also have a distinct archival character to them; the package of 200 episodes of the Sci-Fi TV series *Mystery Science Theatre 3000* contains 135 GB of data—a media archive in its own right. In fact,

since approximately 50 percent of Internet traffic is made up of media files being up- and downloaded in various P2P-systems, the digital domain can be described as one giant media transfer network built upon an organic and rapidly increasing archive of cultural content. The more the online archive is used the bigger it gets; once you're downloading you are always uploading. P2P-protocols work acts exactly the other way around than HTTP-protocols'. The genius of BitTorrent and the like are that while a website is loading slow if there are too many users, P2P-protocols' operate in the opposite way—the more users, the faster the distribution and the more "the archive" grows.

One might, hence, argue that the archival mode of online media is apparent on at least three levels. First, there are the P2P networks (described above), secondly there are the actual media archives on the Web, and thirdly the storage mode is vivid in archival formats of media. The Internet Archive, the World Digital Library or Europeana are examples of online archives where metadata and media material are no longer separated as in analogue media archives. The archival mode is, thus, apparent in the way the computer screen functions as the actual archival interface. In the fancy demo reel for the World Digital Library, for instance, the GUI, the graphical user interface is not only a window to the world, the screen is also an archival interface in time. Texts, images, maps and videos are shown side by side in a multi media display where themes and topics are central rather than the actual media forms.

FIGURE 3: Archival interfaces—screen shot from the concept video of World Digital Library.

Still, archival modes are perhaps most apparent in the third category of new forms of digitally upgraded media, notably television and radio. These new media forms are together with the P2P networks, probably the best examples of how the Web has changed, altered and modified these media in various archival directions. The way these new media forms are presented and promoted also testifies to the apparent storage quality of online media. The archival slogan of the BBC iPlayer for example states: "Making the Unmissable, Unmissable," and in a similar fashion the logo for the Web version of Swedish Television, the so called SVT Play notes: "More than 2.000 hours of TV—whenever you like." SVT Play is, hence, promoted as a distinct archival application, and the same basically goes for public service radio. At the Web page of Swedish Radio you can, for example, find more than 9,000 hours of radio, a neat collection of structured sound in the form of a giant database. Furthermore, music archives built up at sites like Last.fm or Spotify also relates to the same archival mode. Apparently, Last.fm's database contains millions of audio tracks, and Spotify even boasts more than 10 million tracks. The latter is a proprietary P2P and music streaming service that allows instant listening to specific tracks or albums with no buffering delay; music can be browsed by artists or albums as well as by direct searches—all in the form of online archive. In short, Last.fm and Spotify are media archival applications that take advantage of the storage abilities and networkable nature characterizing the current Web.

ACCESS 24

The most important lesson that the Web 2.0 transition has taught, is probably that people are not only connected to the Internet and the Web. They are to an even greater extent connected to each other. Napster and file sharing have paved the way for social networking sites where information is linked, shared and distributed. Worldwide Facebook is, for example, ranked as the sixth most popular site on the Web, with some 275 million regular users. Naturally, Facebook is always accessible, and with the Facebook app on Apple's iPhone a user can do just about anything even while on the move. The bigger picture, however, is that digital usage has led to completely new user patterns in relation to culture in general, and media in particular. Digital user patterns are very different from

previous analogue ones, not the least with regards to the heritage sector. In the mid-1990s, the film archivist Paolo Cherchi Usai wrote an article where he claimed that in traditional film archives some five to ten percent of the holdings are—and will be—used. The rest of the films will remain on their shelf; that is, some ninety percent of preserved films will never be looked at.[12]

FIGURE 4: Analogue vieweing patterns are very different from new digital ones. In a traditional film archive som five to ten percent of the holdings are and will be used—on the Web the opposite holds true.

It is interesting to compare Cherchi Usai's archival estimations to the latest statistics from Comscore regarding online video, figures that were released in March 2009. Almost 80 percent of the total U.S. Internet audience viewed online video, that is, four out of five of every U.S. surfer. Each of these 150 millions video viewers watched a little more than 100 videos equalling some six hours[13]—which of course still is nothing compared to similar figures for television. The point, however, is the way usage pattern changes when moving image databases and archives are accessible at any time on any computer at any place. On YouTube, for example, it is hard to find a video with less than ten views. Thus, almost all videos uploaded are seen by someone, a viewing pattern totally the

opposite to traditional media archives. Basically, the same goes for any type of cultural collection of material on the Web. If users can access material—they will. Unlocking archives and using the generative potential of the Internet and its networks, as put forward in Jonathan Zittrains book, *The Future of the Internet*, seems hence to be an urgent task for heritage institutions.[14]

The media archivist Richard Wright at BBC sometimes refers to this caricature of an archivist as someone who wishes researchers or the public wouldn't come in and disturb and damage the collections. In the digital domain, however, such an attitude is impossible—not the least from a political perspective. Of course, the role of archives is to protect content but the worst way to assure preservation is to deny access, especially at a binary time when "sharing" is a key concept on the Web. Such denial generates no value, and, as Wright has argued in various contexts leaves archives unable to afford to run preservation projects. Protecting content requires public interest—which today comes from digital access.[15]

THE DIGITAL AS DEFAULT

In the binary world everything is plenty; online there is always enough for everybody. The value of digital information is, in fact, hardly measurable, and lack is a word lacking in the digital domain where the cost of distributing information and reproducing it is neglible. As people like Chris Anderson keeps reminding us, there has never been a more competitive market than the Internet, and every day the marginal cost of digital information comes closer to nothing. According to an article in *New York Times* so called "freemium" is becoming the "most popular business model among Web start-ups."[16] In short, freemium is sort of a business model that works by offering basic services for free, while charging for advanced or special features. Slowly these ideas are also heading towards the heritage sector. Rick Prelinger's collection at Internet Archives is, for once, free to use without restrictions, and more than 2,000 film files can be downloaded from the Web. Free has in fact been Prelinger's way of making a business. The Prelinger Archive's "goal remains to collect, preserve, and facilitate access to films of historic significance that haven't been collected elsewhere. Included are films produced by and for many hundreds of important US corporations, nonprofit organizations, trade

associations, community and interest groups, and educational institutions."[17] Letting producers browse and look at footage for free, and then charging them for high-resolution imagery has proven to be a business model more profitable than not providing access.[18]

Within parts of the archival sector there are, however, those who are still in doubt regarding "the digital." In the age of digital reproduction an organization like FIAF for instance, the International Federation of Film Archives, hardly see binary code as the default value for archives and heritage institutions. Digital preservation is as volatile and unsure, as it is expensive—so goes the argument.[19] However, digital media is not ethereal, and storage costs keep dropping, although maintenance and service contracts remain a problem. Furthermore, even virtual reality has a material foundation in the form of nano technological inscriptions on the computer's hard disk. Buildings might collapse and server systems might get flooded, but binary information is nearly always retrievable. Computer forensics teaches that it is more or less impossible to erase a hard drive; every digital inscription leaves a trace—if only at the nano level.

One might suspect that the archival mistrust regarding digital formats has to do with access, which of course is the flip side of every digital activity. Since preservation for most heritage institutions has always been prior to the type of access that comes without saying with digital formats, the current digitization trend forces archives to deal with the issue of opening themselves. Some institutions have, however, embraced the new situation. The Library of Congress, for example, announced during spring 2009 that they would start uploading millions of sound clips and video files onto iTunes and YouTube. The library already offers most of that media material at its own Web site, but the expansion is part of a "broad strategy to 'fish where the fish are,' " according to Matt Raymond, the library's director of communications.[20] And in September 2009 it was also reported that the U.S. Academy of Television Arts & Sciences Foundation would upload its voluminous collection of interviews with TV industry legends. The oral histories of one medium, television, were hence "being made available to the public via another medium, the Internet. The academy founded its preservation arm [. . .] more than a decade ago. Its goal was to record interviews with stars, producers, writers and executives." All in order "to create a digital encyclopedia of TV history," to quote Karen Herman, the director of the archive. Yet, until the archive embarked on a digitization process, its stacks of videotapes "lan-

guished in a temperature-controlled vault," accessible only to researchers who visited the academy. On the Web, however, the material is now present both on a dedicated YouTube channel and at emmytvlegends.org; both sites "allow visitors to browse by people, shows, professions and topics, and flags highlights within the videos."[21]

At least in the U.S., it is the so-called Flickr Commons project that has paved the way for various heritage institutions new warm embrace of the Web. Flickr is one of most popular photo- sharing sites online, or as Luc Sante has stated: "Flickr is to photography what the Pacific Ocean is to water."[22] Flickr is owned by Yahoo and as a hybrid economy makes money through subsriptions and advertisement. The Flickr Commons project, however, is different. Driven together with the Library of Congress, the leading cultural institution exploring Web 2.0 possibilities, it's key goals are firstly to show "hidden treasures in the world's public photography archives, and secondly to show how input and knowledge can help make these collections even richer." The project started with the Library of Congress uploading some 3,000 photographs from the vast Farm Security Administration's photo archive—an archive which contains almost 180.000 photographs from the 1930s depression until World War II. The project got an overwhelmingly positive response from the Flickr community, and it not only brought public awareness to the Library of Congress' existing online collection, but also sparked creative interaction with the images as users helped provide Library curators with new information on photos. The Library of Congress' photos have received more than 15 million views, and the the Flickr Commons project today has some 50 participating heritage institutions.[23]

Then again, not all institutions are as engaged as the U.S. Library of Congress, and one major obstacle regarding "the digital" and the distrust of it within the heritage sector also has to do with the materiality of culture. The materality of the cultural artefact has (nearly) always been more important than the actual stored content at most memory institutions. Basically, this is what museums do; they collect original pieces of art and craft. Copies of theses cultural objects, be they pots or paintings, would not be worthy collecting even though a digital version of Hieronymus Boschs "The Garden of Delight" made up of 1,600 digital photographs (as the one on Google Earth) presents aspects of this painting that a visitor at Prado in Madrid could never see.

As a consequence, if a film for example has been shot on celluloid, the

only way to stay true to its "originality" from a strict film archival perspective is to screen it on celluloid in a projector or in an editing table. Digitizing a feature film and viewing on an iPhone in mpeg4-format is hence perceived as an archival violation. Naturally, every digitization activity involves some kind of media transfer, where copying might lead to loss of information. This is especially the case with various forms of compressed file formats. The problem, however, is that historically there has been a proliferation of various storage formats. Archival holdings of moving images for example are sometimes almost impossible to view because they are fixed in outdated formats, let alone difficult to copy—if one doesn't consider such an activity as unethical (which some film archives actually do).

FIGURE 5: A digital Kane? Within the heritage sector the materality of the cultural artefact is still more important than the stored content.

It goes without saying that as long as the materiality of culture at heritage institutions is perceived as more important than the actual content itself—stored on various forms of carriers, be they paper, tinfoils or celluloid—the archival transition to the digital domain will be utterly slow and remain difficult to steer. Conseptualizing the digital as the archival default does not mean avoiding storage strategies, nor that the materiality of content will be neglected or thrown away. It simply means that for example preserved celluloid films, frozen down and vacuum sealed (which according

to the Swedish Film Institute will make films last for 500 years), do not make sense *at all* in relation to the digital domain. Since resources and funding are always limited within the heritage sector, *access* needs to be the new guiding principle rather than preservation in years to come. The binary paradigm has made analogue preservation strategies so obsolete that in the long run it will become difficult to finance them. The lack of preservational funding in many countries, and the indifferent attitude among politicians and policy makers regarding national heritage, has sometimes been caused by the ALM-sector itself. Heritage institutions in Europe at least, have for years complained that too little money has been spent on them. But the argument has remained the same: give us more money so that we can collect and store things in our stacks and vaults. However, the political impact, the public relations value or the attention of media of such claims often amounts to nothing. Increased funding for putting "films on shelfs" (as I once heard a FIAF representative stating) is hardly the most progressive way forward. Younger users brought up with the Internet will simply not accept that public money in years to come is spent on preservational strategies at heritage institutions that do not involve any form of access to the material.

So whether or not one likes the differences of cultural artifacts in digitized form dissolve into a pulsing stream of bits and bytes, heritage institutions are faced with the fact that the Web and sites like YouTube, Flickr or Wikipedia have become the new archival interfaces. These sites are naturally not archives in a traditional sense—they don't store and preserve material—but in terms of access they have established a radically new archival paradigm. For some archivist within the film heritage sector, the archival mode of online media has become evident predominantly with YouTube's collection of perhaps 200 million videos, making the Internet the world's largest vault for moving image material. Others have stressed the lack of quality and preservational strategies. Kristin Thompson have for example argued that the "celestial multiplex" is a myth, and that there will "never come a time when everything is available [online]." And besides most film "archives are more concerned about getting the money to conserve or restore aging, unique prints than about making them widely available."[24] Sadly, Thompson might be right, but such an attitude has inevitably led to a paradoxical situation. As heritage institutions insist on the importance of traditional and classical archival missions, "they will appear to be less useful, less accommodating, less

relevant, and ultimately less important" than the new archival sites on the Web, not the least from the political perspective regulating funding. As Rick Prelinger has noted, everything that anyone does to bring especially film archives online is from now on always "going to be measured against YouTube's ambiguous legacy."[25] Like Google, YouTube has constantly been distributing itself. Whereas YouTube.com rapidly established itself as the default site for online video, with the consequence that professional content providers lined up for partnership, surfers also encountered YouTube videos everywhere on the Web. The circulation of videos, especially the ability to embed them at other sites, blogs and social networking pages is crucial for understanding the success of YouTube. The site was—and is—both a node and a network.[26]

SOCIAL TAGGING

Wikipedia informs us that social tagging is the "practice and method of collaboratively creating and managing tags to annotate and categorize content."[27] Folksonomy, on the other hand, describes the bottom-up classification systems that emerge from the act of social tagging. The ALM-sector has been hesitant towards letting users provide heritage material with new metadata. This is understandable. However, examples from the Web—predominantly Wikipedia—testify to the wisdom of the crowds. If crowd sourcing can be useful for other domains, there is no reason why the heritage sector should be an exception. The Flickr Commons project is a case in point. The report summary issued by the Library of Congress, "For the Common Good: The Library of Congress Flickr Pilot Project," stated that comments on the provided photographs "have turned out to be very interesting and informative." Furthermore, the project "significantly increased the reach of Library content and demonstrated the many kinds of creative interactions that are possible when people can access collections within their own Web communities. The contribution of additional information to thousands of photographs was invaluable." The displayed photographs had allowed viewers to make reminiscences and share knowledge. "Drawing on personal histories, Flickr members have made connections between the past and the present, including memories of farming practices, grandparents' lives, women's roles in World War II, and the changing landscape of local neighbor-

hoods. Sometimes commenters have identified the precise locations of photos, and [...] offered corrections and additions by identifying locations, events, individuals, and precise dates."[28] Apparently, after verification, social tags and information provided by the community was incorporated into the library catalog. More than 500 records in the original catalogue have been enhanced with new metadata that cites the Flickr Commons project as the source of the information changed or added.

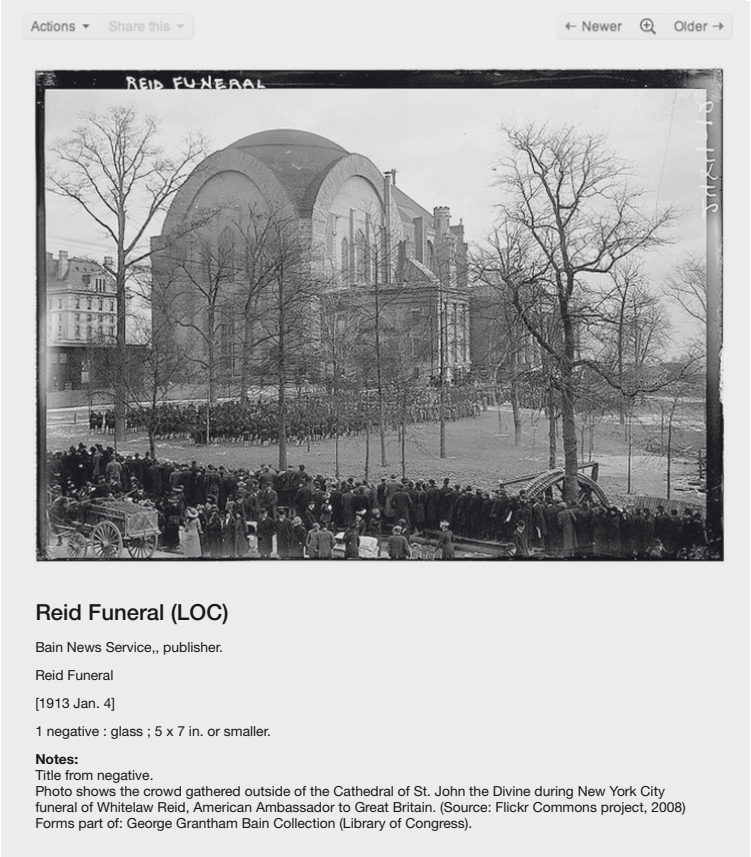

FIGURE 6: The Library of Congress Flickr project has through social tagging led to better data in more than 500 records of the original catalogue.

Regarding social tagging, however, it remains important to note that even if the Web has spurred social interactivity and grassroot' activities, according to the so-called "90-9-1 rule," only a fraction of users do most

of the tagging. In short, the "90-9-1 rule" stipulates that 90 percent of online audiences never interact, nine percent interacts perhaps a handful of times, and only one percent does most interacting. Nevertheless, on a global scale one percent of users might add up to tens of thousands of people. Still, social tagging has often been criticized because of the lack of control, especially in terms of terminology. Without a strict system it is likely to produce unreliable and inconsistent result; errors and contradictions are at times certainly the case. However, one might argue that the Web in general is moving towards an informational space where most data has to be filtered, valued and estimated. In addition, Wikipedia, YouTube and other "information spaces" of course vary a lot. Suffice it to say, they are not either, or. Some information provided by users might indeed be inadequate—and some information excellent. Nowhere else is there such an updated information on digital media, P2P, APIs and the semantic Web as on Wikipedia for example. It all depends on who does the interaction, who contributes and who tags. But heritage institutions should of course like the Library of Congress use the wisdom of the crowd, especially if catalogue information is scarce. The socially tagged information can then be filtered through the institution. In an age when copy and paste are as normal as breathing in front of a computer, heritage institution needs to become aware that they cannot produce all information themselves. A more efficient and updated way of providing apt metadata is to re-use and edit information that already exists on the Web.

DATA MINING

Digitized collections of heritage material can today produce the most amazing things in terms of new knowledge structures. Different ways of data mining, that is the process of extracting more or less hidden patterns from huge amounts of data, has become an increasingly important tool to transform data into information. Data mining is, in short, the process of using computing power to retrieve new techniques for knowledge discovery. There are many nuances to this process, but roughly the steps are three. Firstly, one has to pre-process raw data, secondly mine the data, and finally interpret the results.[29] "More data is better data," Google claims, and in general Google is on its way to make statistical selection old-fashioned. Google already has enough data to make apt statistical

analysis of basically anything with regards to the Web — you don't need to select, you simply mine everything.

However, data mining has also made its way into the heritage sector. An interesting project that gives a vivid impression of what is at stake is the so-called "Digging into Data Challenge", an international grant competition sponsored by four leading research agencies.[30] The idea behind the "Digging into Data challenge" is to answer questions like: "what do you do with a million books, or a million pages of newspaper, or a million photographs of artworks?" That is, "how does the notion of scale affect Humanities and social science research? Now that scholars have access to huge repositories of digitized data—far more than they could read in a lifetime—what does that mean for research?" The concept of data mining indicates that new techniques of large-scale data analysis can allow researchers to discover new relationships—or discrepancies. By performing computations on data sets that are "so large that they can be processed only using computing resources and computational methods developed and made economically affordable within the past few years" new knowledge and information can be deduced.

Within sciences as astronomy the "mining trend" has been going on for years; the SETI@home project, for example, a distributed computing project with more than five million Internet-connected PCs, tries to search for extra-terrestrial intelligence in space. But data mining technology can nowadays also be applied to the large collections of scanned heritage material across the globe. With books, newspapers, journals, films, artworks, and sound recordings being digitized on a massive scale, it is possible to apply data analysis techniques to large collections of diverse cultural heritage resources. Hence the "Digging into Data challenge" ask how these techniques might help scholars use digitized material to ask new questions. The goals of the initiative are: "to promote the development and deployment of innovative research techniques in large-scale data analysis; to foster interdisciplinary collaboration among scholars in the Humanities, social sciences, computer sciences, information sciences, and other fields, around questions of text and data analysis; to promote international collaboration; an to work with data repositories that hold large digital collections to ensure efficient access to these materials for research."[31]

As a process of extracting information from large digital sets of material, data mining does not work in the analogue archive. It is, hence, a process that indicates what one can do, and what can happen once heritage mate-

rial has been digitized. Within the ALM-sector there is sometimes a belief that when material has been digitized, the binary process has come to a halt. But in reality this is where the fun begins. When material exist in binary form, computers can start working on these digital assets deducing information on for example color settings in 10,000 digital impressionistic paintings or the clip rate in 5,000 Swedish newsreels. Scholars in the Humanities always known that quantitative analysis will lead to better results than sample analysis. The best researcher is often the one who is able to go through as much material as possible. But since it has been more or less impossible to deduce patterns out of large sets of cultural objects prior to the usage of computers, the hermeneutical tradition of close qualitative readings have been favored. What data mining suggests, however, is that cultural and art sciences dealing with the past are now able to shift from a hermeneutical perspective to a kind of cultural science of heritage material—or as Daniel J. Cohen stated a few years ago:

> Quantity [can] may make up for a lack of quality. We humanists care about quality; we greatly respect the scholarly editions of texts that grace the well-tended shelves of university research libraries and disdain the simple, threadbare paperback editions that populate the shelves of airport bookstores. The former provides a host of helpful apparatuses, such as a way to check on sources and an index, while the latter merely gives us plain, unembellished text. But the Web has shown what can happen when you aggregate a very large set of merely decent (or even worse) documents. As the size of a collection grows, you can begin to extract information and knowledge from it in ways that are impossible with small collections, even if the quality of individual documents in that giant corpus is relatively poor.[32]

Basically, this is what Google has been doing in its Book Search Project. Another example of actual research in this field is the sort of "cultural analytics" that Lev Manovich has embarked upon. "The explosive growth of cultural content on the web including social media and the digitization efforts by museums, libraries, and companies since the 1990s make possible fundamentally new paradigm for the study of both contemporary and historical cultures," Manovich has stated. One of the reasons behind cultural analytics is that through various data mining sets, cultural objects in digital form as paintings, texts, photographs or films can as data files *per se* produce new forms of metadata. In other words, from one single nu-

merical file a lot of data can be extracted — and from thousands of them entire data sets. Computer-based techniques for quantitative analysis and interactive visualization can, thus, be used to analyze patterns in massive cultural sets. Manovich is the leading figure behind the so called "software studies initiative" at the University of California, San Diego, with a firm belief "that a systematic use of large-scale computational analysis and interactive visualization of cultural data sets and data streams will become a major trend in cultural criticism and culture industries in the coming decades." At the core of the initiative lies the question what will actually happen when scholars start using interactive visualizations of large data sets as a standard tool in their work (like many scientists do already). Perhaps new cultural disciplines will "emerge out of the use of interactive visualization and data analysis of large cultural data sets?"[33]

FIGURE 7: Cultural analytics of Dziga Vertov's *The Man with the Movie Camera* (1929). In digitised form one can mine and run the film through a sophisticated computer grid and deduce average shot length and other forms of cinemetrics.

THINK DISTRIBUTED—OR STORAGE AS NETWORK

During the last three year the so-called cloud computing trend has transformed parts of the Web into a platform of distributed storage. In short, cloud computing defines a new kind of infrastructure for personified information, which no longer exists and resides locally on one's own computer, but online in the Internet's network. On a theoretical level the new digital cloud can be seen as something that substantially changed how we comprehend view the computer as a machine. Today it seems clear that our computers cannot be understood as isolated and separate units. Concurrent with development of the Web, it has become impossible to differentiate one's own computer from the network it has become an unmistakable part of. Using the small free application Dropbox, for example, one can easily (by placing a file in a folder) use the Web as both a storage place and a file server and network between different computers and mobile devices.

What this shift implies for the heritage sector is still hard to say. However, as trust and reliability of online technology increases, institutions do have to acknowledge the possibilities that networked binary technologies offer. If the Web's network of networks of computers and servers can be perceived as one gigantic information processing machine, which through sophisticated and super-fast communication protocols share bits of data and strings of code, real archival centralization through new binary distributive methods is for example a possibility. It has been argued that Google is a post-media company that operates and thinks in distributed ways. Bits of Google are all over the Web, and even though there is a google.com-page—it too is but a node belonging to the company network. As is well known, the main purpose behind the Arpanet/Internet, set up in the 1960s during the cold war, was its decentralized character. If one part of the net was damaged, it would not mean that other areas stopped working. Within the heritage sector archivist are also afraid that their digital files will become corrupt and damaged. However, by using the networkable nature of new digital media networked preservation is an area with a number of interesting projects. The LOCKSS-project, that is "Lots of Copies Keep Stuff Safe", is, for example, an international community initiative that provides libraries with digital preservation tools in the form of open-source software, so that they can easily collect and preserve their own copies of authorized e-content. Furthermore, libraries

cooperate in a P2P-network to ensure preservation of that content. By way of cryptographic hash functions, a damaged copy of a book or a journal can within LOCKSS easily be repaired from other peers in the network. In short, LOCKSS uses file sharing as a preservational model.[34]

Another example based on the same premise of distributed storage is the "Chronopolis. Preserving Our Digital Heritage"-project, a multimember partnership run by the Library of Congress at the San Diego Supercomputer Centre. A key goal of the Chronopolis project is to provide cross-domain collection sharing for long-term preservation. Hence, it is part of a new breed of distributed digital preservation programs that tries to use the networkable nature of new digital technology. Long-term digital preservation is still an issue that most heritage institutions struggle with. The Chronopolis project addresses this critical problem by providing "a comprehensive model for the cyberinfrastructure of collection management, in which preserved intellectual capital is easily accessible, and research results, education material, and new knowledge can be incorporated smoothly over the long term." By integrating digital library resources, various data grids and persistent archive technologies, Chronopolis seems to be on its way to create a kind of trusted "preservation environments that span academic institutions and research projects, with the goal of long-term collection management, preservation, and knowledge generation."[35]

One of the more interesting aspects of the Chronopolis project is its cross-cultural approach between both the Humanities and the natural sciences, and between commercial and government entities. In recent years, such institutions and companies have all produced reports and funded investigative efforts on aspects of the problems of data management and long-term digital preservation. What the Chronoplis project suggest is that the efforts of institutions as diverse as libraries and museums, science and engineering funding agencies, as well as supercomputing centers, should all be seen a complementary, "although these institutions may not have long histories of collaboration and have seemingly focused on very different disciplinary activities in the past."[36] On the Web as well as on the Internet, the network is everything, and new storage technologies of course tries to use and "think" preservation through the digital technology itself. There are, in fact, also various new sets of smaller distributive storage technologies. When one preserves a digital object in these systems it is, basically split up into chunks stored on various machines spread across a network in the manner of file sharing. These

bits of data can then be replicated or encoded for redundancy — hence the storage capacity can be very reliable. The Open Source project "Celeste" is, for example, a highly available distributed, P2P- data storage system.[37]

CONCLUSION—THE ARCHIVE AS PLATFORM

Media critic Kevin Kelly has made the claim that sharing seems to be key to success in the digital domain. At the "Web 2.0 Summit" conference in San Francisco 2008, Kelly pointed out that because our media are converging, there would soon be only one common cultural platform. Everything is run by the same kind of shared Web-based machine independent of device. In his talk, Kelly stressed that in the future, three overall trajectories will probably characterize the Web: a move up into the "digital cloud," a move down into databases, and a move towards a kind of general sharing. According to him, information that is not accessible—will cease to exist.[38] Open source, various "mash-up" technologies and services, as well as open APIs are already progressing in this manner. One might even argue that if Gutenberg's movable pieces of type were the modules on which the art of printing rested, then almost analogous divisible binary program modules will—if shared—constitute the foundation of the information landscape of the future.

Today, a circulating binary flow of texts, images, sound and video characterizes the Web. Open APIs makes it possible for programmers to constantly develop new interfaces. Sharing metadata through open source and open APIs is also a way for the ALM-sector to increase a dynamic growth of new binary heritage services. Following Google, many websites have exposed their APIs and made them available to external developers, and even though there are numerous initiatives within the heritage sector, particularly concerning libraries, there is a general lack of data within the heritage domain. If data was more frequently re-used and shared, more sophisticated system could be built and creative mash-ups constructed, which in turn would mean that the collected data would reach a wider audience.

This article has dealt with different ways in which heritage institutions should use and think about digital technologies in general, and the Web in particular. My last proposition suggests that the ALM-sector

should upgrade itself, and perceive itself a sort of binary platform within the digital domain. The notion of platform, in fact, sums up what the article has tried to address, namely an increased focus on access and distribution, interactivity and sharing. The term "platform" can of course mean many things: a computing platform often describes some sort of software framework, and a political platform is a list of principles held by a political party. Yet as Jeff Jarvis has stated, "a platform enables. It helps others build value. [. . .] Networks are built atop platforms."[39] A number of popular social networking sites like to perceive themselves as more or less "empty" platforms to be filled by the communities using them. YouTube, for example, presents and views itself as a platform — and not as a regular media distributor (especially when copyright issues are at hand). For sure, there is a distinct politics of contemporary online platforms. However, I would argue that as a digital platform heritage institutions could better fulfill their basic democratic functions. If institutions would establish themselves on the Web and be open and collaborative with data and digital material, rather than claiming content as their own assets (like the National Portrait Gallery), users will naturally start adding value to such platforms.

ENDNOTES

1. "U.K. National Portrait Gallery threatens U.S. citizen with legal action over Wikimedia images" 14 July 2009—http://en.wikinews.org/wiki/U.K._National_ Portrait_Gallery_threatens_U.S._citizen_with_legal_action_over_Wikimedia_images [last checked 15 September 2009]. For a discussion, see also Wikimedia Commons "User:Dcoetzee/NPG legal threat"—http://commons. wikimedia.org/wiki/User:Dcoetzee/NPG_legal_threat[last checked 15 September 2009].
2. Cory Doctorow, "UK National Portrait Gallery threatens Wikipedia over scans of its public domain art" *boingboing* 20 July 2009—http://www.boingboing.net/2009/07/20/uk-national-portrait. html [last checked 15 September 2009].
3. Eric Felten, "Who's Art Is It, Anyway?" *Wall Street Journal,* 30 July 2009—http://online.wsj.com/article/SB1000.html [last checked 15 September 2009]. For a discussion, see also Kenneth Hamma, "Public Domain Art in an Age of Easier Mechanical Reproducibility" *D-Lib Magazine,* No. 11 (2005)— http://

www.dli b.org/dlib/november05/hamma/11hamma.html [last checked 15 September 2009].

4. Astrid Verhausen, "Mass Digitisation by Libraries: Issues concerning Organisation, Quality and Efficiency" *Liber Quaterly*, No. 1 (2008)—http://webdoc.sub.gwdg.de/edoc/aw/liber/lq-1-08/ar ticle4.pdf [last checked 15 September 2009].
5. The funding provided from Google to various libraries taken part in the Book Search project is confidential. At a meeting in Munich in March 2009 with Klaus Ceynowa, the deputy Director General of the State Library of Bavaria, confessed however that around € 50 million is what the library receives in funding.
6. "Europe's Digital Library doubles in size but also shows EU's lack of common web copyright solution", press release 28 August 2009—http://europa.eu/rapid/pressReleasesAction.do?reference=I P/09/1257&format [last checked 15 September 2009].
7. Lev Manovich, *Software Takes Command* (2008)—http://software studies.com/softbook/manovich_softbook_11_20_2008.pdf, p. 19. [last checked 15 September 2009].
8. Ibid., 18.
9. Lawrence Lessig, *Remix: Making Art and Commerce Thrive in the Hybrid Economy* (London: Penguin Press, 2008).
10. Geert Lovink, "The Art of Watching Databases: Introduction to the Video Vortex Reader," *Video Vortex Reader: Responses to YouTube* Geert Lovink & Sabine Niederer (eds.) (Amsterdam: Institute of Network Cultures, 2008), 9—http://network cultures.org/wpmu/portal/publications/inc-readers/videovortex/ [last checked 15 September 2009].
11. Lev Manovich, *The Language of New Media* (Cambridge, Mass.: MIT Press, 2001).
12. Paolo Cherchi Usai, "Äh, det är ju bara journalfilm: Varför ses inte 95% av museernas bestånd" *Aura* No. 3–4 (1997).
13. "YouTube Surpasses 100 Million U.S. Viewers for the First Time", Comscore press release 4 March 2009—http://comsc ore.com/index.php//Press_Events/Press_Releases/2009/3/YouTube_Surpasses_100_Million_US_Viewers [last checked 15 September 2009].
14. For a discussion, see Jonathan Zittrain, *The Future of the Internet* (New Haven: Yale University Press, 2008).
15. Richard Wright has made similar claims on various media archival events during the last decade. For a discussion, see for example the panel "Institu-

tional Perspectives on Storage" at the MIT "Media in transition 6" conference in Boston, April 2009—http://mitworld.mit.edu/video/681/ [last checked 15 September 2009].

16. Clair Cain Miller, "Ad Revenue on the Web? No Sure Bet" *New York Times,* 24 May 2009—http://www.nytimes.com/2009/ 05/25/technology/start-ups/25startup.html [last checked 15 September 2009].
17. For a discussion around the Prelinger Archives, see http://www.archive.org/details/prelinger [last checked 15 September 2009].
18. Matthew B. Kirschenbaum, *Mechanisms: New media and the Forensic Imagination* (Cambridge, Mass.: MIT Press, 2008).
19. For a discussion see http://www.fiafnet.org as well as Paolo Cherchi Usai *et al.*, *Film Curatorship: Archives, Museums, and the Digital Marketplace* (Vienna: Synema, 2008).
20. For a discussion see, Grant Gross, "Library of Congress embraces YouTube, iTunes", 27 March 2009, *IDg News Service*—http://www.networkworld.com/news/2009/032709-library-of-congress-embraces-youtube.html [last checked 15 September 2009].
21. See, Brian Stelter, "Interviews With Legends of Television Hit the Web" 13 September 2009—http://www.nytimes.com/2009/09 /14/ business/media/14archive.html [last checked 15 September 2009].
22. Sante is quoted from, Noam Cohen, "Historical photos in web archives gain vivid new lives", *New York Times,* 19 January 2009—http://www.nytimes.com/2009/01/19/technology/internet/19link.html [last checked 15 September 2009].
23. For a discussion, see—http://www.flickr.com/commons [last checked 15 September 2009].
24. Kristin Thompson, "The Celestial Multiplex", 27 March 2007—blog post at http://www.davidbordwell.net/blog/?p=595 [last checked 15 September 2009].
25. Rick Prelinger, "The Apperance of Archives" Pelle Snickars & Patrick Vonderau *The YouTube Reader* (eds.) (Stockholm: KB, 2009), 272.
26. For a discussion, see *The YouTube Reader* (2009).
27. See—http://en.wikipedia.org/wiki/Social_tagging [last checked 15 September 2009].
28. Michelle Springer *et al.*,"For the Common Good: The Library of Congress Flickr Pilot Project" (2008)—http://www.loc.gov/rr/pri nt/flickr_report_final.pdf [last checked 15 September 2009].
29. For information on "data mining", see Wikipedia—http://en.wikipedia.org/wiki/Data_mining [last checked 15 September 2009].

30. For more information, see—http://www.diggingintodata.org/ [last checked 15 September 2009].
31. Ibid.
32. Daniel J. Cohen, "From Babel to Knowledge. Data Mining Large Digital Collections", *D-Lib Magazine* no. 3, 2006—http://www.dlib.org/dlib/march06/cohen/03cohen.htm [last checked 15 September 2009].
33. Lev Manovich, "Cultural Analytics"—http://lab.software studies.com/2008/09/cultural-analytics.html [last checked 15 September 2009].
34. For a discussion, see—http://www.lockss.org/lockss/Home [last checked 15 September 2009].
35. For a discussion, see—http://chronopolis.sdsc.edu/index.html [last checked 15 September 2009].
36. Christopher Jordan *et al.*, "Encouraging Cyberinfrastructure Collaboration for Digital Preservation" (2008)—http://chronopolis. sdsc.edu/assets/docs/39_Jordan.pdf [last checked 15 September 2009].
37. See—http://www.opensolaris.org/os/project/celeste/ [last checked 15 September 2009].
38. Kevin Kelly's talk can be found at—http://www.youtube.com/watch?v=1So-S36pMo4 [last checked 15 September 2009].
39. Jeff Jarvis, *What Would Google Do?* (New York: HarperCollins, 2009), 32.