Keynote Address: Sport Archives and Digitalization

Amit Solomon
Fahrenheit 452, India

In 1971, a young man was given an operator's account with 100 million dollars worth of computer time in it by the operators of the Xerox Sigma V mainframe at the Materials Research Lab at the University of Illinois.

This was totally serendipitous, as it turned out that two of the four-operator crew happened to be his best friend and the best friend of his brother. At the time there was more computer time than people knew what to do with, and those operators were encouraged to do whatever they wanted with that fortune in "spare time" in the hopes they would learn more for their job proficiency.

At any rate, this young man decided there was nothing he could do, in the way of "normal computing," that would repay the huge value of the computer time he had been given ... so he had to create 100 million dollars worth of value in some other manner.

An hour and 47 minutes later, he announced that the greatest value created by computers would not be computing, but would be the storage, retrieval, and searching of what was stored in our libraries.

He then proceeded to type in the "Declaration of Independence" and tried to send it to everyone on the networks ... which can only be described today as a not so narrow miss at creating an early version of what was later called the "Internet Virus." A friendly dissuasion from this yielded the first posting of a document in electronic text.

The name of the young man who posted that first digitized document? Michael Hart – the founder of Project Gutenberg.

We have come a long way since then but Michael Hart's vision still holds true – that the greatest value created by computers would be the storage, retrieval, and searching of what is stored in our libraries.

I'm a partner in a firm that offers digitization services to publishers and libraries amongst other services. We began offering these solutions – that go under the name Fahrenheit 452 - around five years back and five years of digital conversion work, day in and day out has provided me with some important insights that I would like to share with you today. There are a number of things I could have shared with you that are of a technical nature or given a report on where digitization is headed but I've decided to share with you something far more fundamental in nature.

I have seen digitization at very close quarters. I have lurked in online forums dedicated to digitization and listened in on people venting their feelings on digitization, people asking questions, people offering suggestions and predictions. Taken part in online discussions that turned into extremely long multi-threaded debates. I have read about - and then worked with - different technologies, many of those were touted as the next big thing but came up short in practice.

Two and a half years back, we started working with the Amateur Athletic Foundation to create digital versions of various Olympic Reports, Sport Magazines and Journals. That gave me my first experience with digitizing a sport archive. And today I stand here in a Sports Congress – addressing people who are eager to see sport information reach the furthest corners of the world.

I want to start by telling you a story that you've probably heard before - or read. Now why would I want to re-tell a well-known story? Because I believe it’s the best way to illustrate the thoughts I want to share with you today.

The story is told of a Persian by the name of Al Hafed. Al Hafed owned a very large farm with orchards, grain fields and gardens. He was a contented and wealthy man.

One day an old priest visited Al Hafed and while they were talking, the priest happened to remark that a diamond is the last and highest of God's mineral creations, as a woman is the last and highest of God's animal creations. And that is the reason why the two have such a liking for each other. The priest continued and told Al Hafed that if he had a handful of diamonds he could purchase a whole country, and with a mine of diamonds

he could place his children upon thrones through the influence of their great wealth.

Al Hafed heard all about diamonds and how much they were worth, and he could not sleep that night. The talk with the priest had made him discontented. He said to himself: "I want a mine of diamonds!" So he lay awake all night, and early in the morning sought out the priest.

He awoke that priest out of his dreams and said to him, "Will you tell me where I can find diamonds?" "Well," said the priest, "if you will find a river that runs over white sand between high mountains, in those sands you will always see diamonds." "Do you really believe that there is such a river?" "Plenty of them; all you have to do is just go and find them, then you have them." Al Hafed said, "I will go." So he sold his farm, left his family in charge of a neighbor, and away he went in search of diamonds.

He began very properly at the Mountains of the Moon but found nothing. Afterwards he went around into Palestine but still no luck. He then wandered on into Europe, and at last, when his money was all spent, and he was in rags, wretchedness and poverty, he stood on the shore of that bay in Barcelona, Spain, when a tidal wave came rolling in through the Pillars of Hercules and the poor, afflicted, suffering man could not resist the awful temptation to cast himself into that incoming tide, and he sank beneath its foaming crest, never to rise in this life again.

The story does not end there.

One day Al Hafed's successor led his camel out into the garden to drink, and as that camel put its nose down into the clear water of the garden brook Al Hafed's successor noticed a curious flash of light from the sands of the shallow stream, and reaching in he pulled out a black stone having an eye of light that reflected all the colors of the rainbow, and he took that curious pebble into the house and left it on the mantel, then went on his way and forgot all about it.

A few days after that, this same old priest who told Al Hafed how diamonds were made, came in to visit his successor, when he saw that flash of light from the mantel. He rushed up and said, "Here is a diamond! Has Al Hafed returned?" "No; Al Hafed has not returned and that is not a diamond;

that is nothing but a stone; we found it right out here in our garden." "But I know a diamond when I see it," said he; "that is a diamond!"

Then together they rushed to the garden and stirred up the white sands with their fingers and found other more beautiful, more valuable diamonds than the first, and thus, according to legend, were discovered the diamond mines of Golconda, the most magnificent diamond mines in all the history of mankind.

There are a large number of libraries - sport archives included - that are seeing digitization initiatives succeeding at a number of places and are saying to themselves - "We, too, want to build a digital library." And like Al Hafed, they are looking here and there, reading this and that, worried about technology that might soon become obsolete, worrying about the finances, and hoping - hoping all the time - that a technology or a process may soon be developed that would let them begin digitizing their entire collection.

Al Hafed, had he stayed home instead of wandering through strange lands, and dug in his own backyard, would have discovered wealth beyond imagination. Similarly, there are easy - and low cost - ways to begin digitizing your sport archive without waiting for outside help.

But before you start digitizing, you need to get your priorities right. What priorities am I talking about? I'm talking about the reason you want to get started with digitization. What is your first priority - preservation of your collections or increased access to your collections? In a survey done by the Institute of Museum and Library Services in the United States in 2002, public libraries, unlike museums, prioritized preservation of material over increased access, which I feel is a misplaced priority.

Talking of priorities reminds me of this story about this guy who was at the final of the Soccer World Cup where Germany and Brazil were squaring off against each other. He had one of the best seats in the stadium that was completely jam packed with spectators. But this man was surprised to see the seat next to him lie vacant for the entire first half.

At half time, he turned to the man sitting on the other side of the vacant seat and said - "I can't believe some fool would book a seat for the finals and then be a no-show."

The other guy slowly turned towards him and replied - "I'm the one who booked this seat for my wife. Unfortunately, she's dead."

This threw the guy off completely and he mumbled his apologies. But as you know you can't keep a passionate soccer fan down for long, so he turned once again to this man and said - "You know it's pretty sad what happened but couldn't you have passed on her ticket to some friend or relative instead of letting it go waste?"

The bereaved man answered - "I would have but they're all attending her funeral."

Yes, we all have our priorities. But lets get them right. Increased access is as important as preservation, if not more. Sport archives are – in a sense – better placed than other libraries because there isn't that much material that is fragile that it needs urgent preservation. Most of the material in the sport archives is from the last century unlike centuries old material in other large libraries and archives.

And even when the material is old and brittle, digitization would not only provide greater access to it but also indirectly help preserve it, as direct physical access to the brittle material could then be limited.

If you start digitizing because you want to first and foremost preserve your collection, you'll be in for disappointment because right now, despite digitization projects mushrooming all over the world, there are no set standards that have evolved. There are no standards for data encoding or data format and many may need to migrate their digital collections in the future when a standard is agreed upon.

People think that computer files are far safer than paper and do not degrade as quickly, but this is far from true. Computer applications and file formats become obsolete in a matter of years. This happens even more quickly with documents saved in proprietary formats such as Microsoft's .doc format.

Archives that have the necessary funds readily available can afford to keep migrating their digital collections to newer versions but what should an archive with limited funding do? Should they keep waiting for a standard to evolve and then start digitizing? Or should they – for the sake of increased

access – adopt a certain technology and hope like crazy that either the technology won't become obsolete soon or that large grants would keep coming in perpetuity? And what about those archives that have no immediate funding available for digitization? Should they just keep waiting for that elusive grant to come by?

When we started out with our digitization service in 2000, it was just my wife, Leena – also my business partner – and I, who were the owners as well as the employees. Till that time, we were a tiny graphic design studio. In terms of resources, we had just one computer, one scanner, PageMaker software and an old version of a basic OCR software package.

Our first project was for Globe Pequot, publishers of the travel series called the Insider's Guides, and I remember scanning and proof-reading the pages at night while Leena worked on the page layouts during the day. Our OCR package wouldn't maintain layout so the layout had to be done manually.

That was the time when we really learned about how to make do with what we had. Because we couldn't afford more expensive software, we trawled the web for freeware that would do part of the job. The going was laborious – and painful at times – but the result was probably even better than what a top line commercial automated software could have given us.

And while we have grown by leaps and bounds and are now a group of twelve, we still keep looking for the best, affordable options that are available that would give us the best results possible.

In the Amateur Athletic Foundation's digitization project, we have been delivering the output as PDF Normal i.e. PDF Image with searchable text. The AAF wanted to digitally reproduce documents in a format that looked as close to the original as possible, was fully searchable, was small in size and was universally known and available – and PDF Normal fit the bill perfectly.

And we haven't just been producing replicas of the original paper documents but also making value additions such as cleaning up and enhancing the look of the photographs and illustrations in early twentieth century magazines. But this is an expensive format to work in and most archives cannot afford it.

So what's the best option for archives with very limited or no funding for digitization? The best option is to scan the selected book or journal and upload the image to your server for viewing on your website.

I can imagine what's going on in the minds of some of you present here. Just scan and put the images up on the website? Wouldn't high-resolution scans that are clear enough to read be big and cumbersome to download? Wouldn't storage and delivery costs increase dramatically?

Not necessarily. Here's a simple and low cost example of how to start digitizing and disseminating sport information quickly.

In terms of hardware all you need is a computer and a scanner – which most libraries already have. You don't have to buy those top of the line ten or fifteen thousand dollar scanners, a regular A4 scanner capable of scanning up to 400 dpi would be good enough – and that should come under USD 200. For another 100 USD, you get an auto-feeder that would feed in cut pages of a book or journal into the scanner and will save you a lot of time.

I'm assuming you have a computer and a website so I won't discuss costs here – but even those are pretty inexpensive these days.

The problem with scanning a book and just putting the images online is that if you increase the resolution of the page for easy viewing, you increase the file size of the image. That results in slower download times for visitors as well as placing a strain on your available bandwidth. Plus the costs of storage increase as well.

Well, that problem just went away. Actually, it went away quite a while back but it seems very few archives know about it. The solution is a format that's pronounced Déjà vu and written DjVu. It's a format that was created by researchers at AT&T and later sold to a company called Lizard Tech.

Let me give you an example that shows the superiority of DjVu over other formats for digitization purposes.

A 400 DPI Color Scan of an 8.5" X 11" page in uncompressed TIFF format is 44000 KB. In JPEG and PDF (Image Only) format, the size comes down to 1300 KB.

Convert it into DjVu and what do you get? A measly 56 KB. Yes, a 400 DPI color scan of an 8.5" X 11" page is only 56 KB in size.

Lo and behold, your bandwidth problems are solved. But that's not all. You can OCR the image and proof read the text, and place it behind the image as hidden text that is fully searchable. Of course, people will have to download the free DjVu reader and install it on their computers just like you have to download the Acrobat Reader.

Document Express, the software that creates DjVu files costs just under USD 400 but there is a free, open source version out as well now.

If you can't afford to have someone proof read your text, forget about it. Just place the images on your website. If you don't have the funds, it's OK to have a digitized version that has no additional benefits over the printed book. Its great to be able to have full-text search, to have metadata, to have digitally bookmarked documents but what if you can't afford it? What if you can't afford to outsource your work to service companies like mine? The least you can do – if you're seriously interested in digitizing your collection - is to just scan your material, convert it into DjVu and put it on your website. Someone who must otherwise travel to a physical location to access a particular journal would be really happy to find easily accessible scans of the journal on the Internet.

As Michael Keller, university librarian and publisher of the Stanford University Press and the HighWire Press very eloquently put it – "There is obviously a huge amount of information on the web. However, information is not quite a generic commodity: Having millions of pages available online is of no immediate value if the information you need is represented only in a book on a shelf to which you do not have access."

If you're worried that people won't be able to find your collection on the web because the text is not readable by a search engine, create a summary and put that text on your site. The search engines will index that. It's not as efficient but it'll do the job.

Store your scans in JPEG format for future reuse. The JPEG format has been there for many years now and I expect it to be around for many, many more years. So when you get more funding for your digitization efforts or when a common standard evolves, you can convert these scans into a format that is much more versatile and also finds universal acceptance.

Remember, you already have whatever it takes to get started on the road to digitization. You don't need more funds - the lack of money has never stood in the way of a strong purpose.

Of course, the funding issues of at least five libraries have been put on the backburner ever since Google announced its digitization project.

Google's primary goal aims at out-of-print material, whether public domain or in copyright. Google maintains that it is meeting library copyright standards. Participants will receive no financial compensation from Google, but the massive digitization project will also cost them nothing. Each library in the program will receive digital copies of the books it has contributed, which it can then use to enhance service to its own patrons.

Participants in the program are the libraries of Harvard, Stanford, the University of Michigan, and Oxford University, as well as the New York Public Library. The result of the multiple-year project would be an online digital library of what could number as many as 30 million volumes. Google hasn't promised a particular number and I don't think the libraries are insisting on a number considering Google is doing it all for free. Of course, it's going to cost Google millions of dollars – they haven't given a number - but they are absorbing all the costs.

Like the other announcements from universities and libraries, Harvard viewed the program as creating an important public good and serving the world. Harvard president Lawrence H. Summers stated: "Harvard has the greatest university library in the world. If this experiment is successful, we have the potential to provide the world's greatest system for dissemination as well."

The Google digitization project - part of the Google Print program - is an amazing development and I look forward to accessing the world's largest virtual library. Google states on its website: "Google Print is our contribution to a diverse body of digital library developments. What we are

doing is not intended to replace or discourage funding for the efforts of others working to digitize library collections. We hope that our entry into this arena will attract needed attention to digital library initiatives worldwide."

While I'm hoping that will happen, I also fear that some libraries may do the opposite.

My fear is that digitization efforts may either be postponed or slowed down in the hope that the Google digitization project would eventually include other libraries and that the other big search engines like Yahoo and MSN may also start such digitization initiatives.

The Google Digitization Project is not the first endeavor of such magnitude. Project Gutenberg, that is basically supported by individual volunteers and the Internet Archive, which is a collaboration between organizations worldwide – are two such initiatives. Google Digitization project is different from the other initiatives as this time around a single, large corporation is undertaking something of this magnitude.

But why shouldn't you postpone or suspend your digitization programs? After all, you'll save money that would have been spent on digitization and instead spend it on improving and growing your current collections? What's wrong with this mother-of-all-free-lunches? There's nothing wrong – if you're invited to the lunch. But if you're not, don't stop eating in the hope that one day you will be invited.

Google is currently running a test project with a very small number of books. Once it straightens out the kinks, it will take quite a few years for it to complete digitizing the vast number of books. 30 million volumes is a huge, huge number. So don't expect Google to look towards any other libraries before that.

Also, realize that Google is not undertaking this only as a huge philanthropic exercise. I'm sure Google means well and is truly interested in building this vast storehouse of knowledge for the benefit of the entire world. But Google is a publicly traded company and its shareholders – like any other shareholders – would not like to see the company sink millions of dollars in this project and recover nothing. So while Google is not charging the libraries, and it hasn't stated that it would be profiting – if at all – from

this project, the fact of the matter is that it's putting in a huge amount of money into the project and it would be justified to expect commensurate returns. But those returns may take some time coming – and so the next phase of its project involving other libraries in all probability will start many years hence.

I also don't see Yahoo or MSN or any other large corporations jumping into the fray right now and signing up other libraries. I believe they would all be watching this project very closely right now and waiting to see if Google burns its fingers or not. And only once they see positive results – and by that I mean profits – would they also enter the arena. So don't hold your breath waiting for them.

One way that the other libraries would benefit from this project is that with millions of public domain books getting digitized, the other libraries may have fewer volumes to digitize, as a substantial number of public domain holdings would be common.

Another thing that libraries – especially Sport Archives – can do is to focus on digitizing those books and journals that are only available with that particular Sport Archive, and not with any of the big libraries associated with the Google Project. The creation of unique collections would help such Sport Archives to not just find funding quickly but would also help them to build and grow their own niche audience.

While it would be great if corporations would pick up the tabs for all your digitization projects, its not going to happen on a worldwide scale anytime in the near future. And for the benefit of sport researchers – and even plain sport enthusiasts – you must continue the race. When such a thing does happen – embrace it with open arms – and then you can rest. But not until then.

If you need more funds, start thinking more creatively. The first thing every library thinks of when looking for additional finances is getting a grant. Grants are good but there are few grants going around for the large number of proposed digitization projects. Grants are good and there's nothing wrong in creating a great proposal and applying to any foundation that might be remotely interested in awarding you a grant but don't stop there. Don't put all your eggs in one basket.

Think out of the box. Contact famous - and rich - sports personalities to make a contribution towards your project. A lot of them would love to support the sport that has given them so much. If they decline or even hesitate, ask if they could speak to their sponsors. Ask for a letter and follow up with the sponsor. Companies tend to listen better when it's the star athlete they're sponsoring who's doing the asking.

Also, while you're looking for funding, it's good to partner with other archives that have similar interests and pool your resources. A joint digitization initiative will not only be more fruitful because of the synergy that is created but will also help you find funding faster.

If you don't have the funds, you can't hire more staff. But you can make do with even a very limited staff. Let me briefly tell you about a specialized digital archive that was begun by a single dedicated person and has now really grown with the support of hundreds of volunteers. No, it's not Project Gutenberg, although that also started and grew that way. This archive is known as the Christian Classics Ethereal Library and can be found online at www.ccel.org.

Harry Plantinga, a Professor of Computer Science at Calvin College, Michigan, US, started this collection by single-handedly scanning, proofreading and then uploading to the Internet the famous book "The Imitation of Christ". Soon he was looking in used bookstores for Christian books that were in the public domain.

In his own words – "The love of those old classic Christian books, the deep appreciation for their continuing value, and a desire to "pay back the net" for benefits accrued motivated me to continue to scan books and put them on the Internet--first on an FTP site, and then on the World Wide Web, when it arrived on the scene."[5]

By then, people had discovered the site and volunteers helped put more books online and by 1997 CCEL had over 200 of the most important books in English Christian literature. Right now, a cursory glance suggests over 500 titles and many more in the pipeline. By 2001, their server was logging about a quarter of a million "hits" per day from a quarter of a million distinct users per month, providing about 250 GB of information in a month. That's equivalent to about a quarter of a million books per month.

If someone is interested in volunteering by scanning, proofreading, or markup, they can contact the volunteer coordinator. Another way to help is to work on a digital facsimile edition: images of pages of books are put on line and a database-driven system allows you to correct pages one at a time. You can even see what people are working on.

Those who have some knowledge of XML are encouraged to markup the documents into ThML. The CCEL uses Theological Markup Language to represent documents. Think of it as HTML with added support for scripture references, indexes, tables of contents, bibliographic information, and other digital library needs. The information goes into a database for automatic generation of index pages. ThML documents are automatically converted to HTML, HTML zip files, text, Open EBook, Palm DocBook, and Microsoft Reader format for convenient use.

They also have quite a collection of MP3 audio files -- essentially books on tape that you can download to your computer or MP3 player. Again – volunteers have done most of the work.

So if staff is a problem, look at building a team of volunteers. There are enough researchers and even non-researchers who love sports, who'd be willing to lend a hand in building your digital archive.

The going might be slow – and it may not always be steady but at least you'll be off from the starting blocks and covering some distance. And until you move – even slowly - you are not going to build any momentum. Once you have the momentum, you'll be able to reach many milestones.

Conclusion

I'd like to conclude by narrating an incident from the 1976 Olympics.

In the 1976 summer Olympic games, the men's gymnastic competition captured the attention of the world. With the crowd roaring in the background, Japan's Shun Fujimoto landed a perfect triple-somersault twist dismount from the rings to clinch the gold medal in team gymnastics.

With his face contorted in pain and his teammates holding their breath, Fujimoto followed a near flawless routine by achieving a stunning

and perfect landing – on a broken right knee. It was an extraordinary display of courage and commitment.

Interviewed later about the win, Fujimoto revealed that even though he had injured his knee during the earlier floor exercise, it became apparent as the competition continued that the team gold medal would be decided by the ring apparatus – his strongest event. “The pain shot through me like a knife,” he said. “It brought tears to my eyes. But now I have a gold medal and the pain is gone.”

Continuing with your digitizing efforts – even when resources are low and when nay sayers abound can be frustrating and painful. But when your efforts enable researchers and sport enthusiasts around the world to access information that would normally be inaccessible to them, all that pain and trouble will be forgotten as you’ll be honored with a gold medal of heartfelt appreciation.

Lets go for gold.

Thank you and God Bless.