

The societal and commercial potential of Digital Preservation (Redacted)

Authors:

Didier da Costa (Intel) & Mike Nolan (Intel)

Contributors to the original deliverable:

Angela Dappert (DPC), Gregor Heinrich (iPharro), Daniel Draws (SQS), John Thomson (CMS),
Wasif Gilani (SAP), Mark Guttenbrunner (SBA), Daniel Draws (SQS), José Barateiro (LNEC),
Jorge Gomes (LIP), Ricardo Vieira (INESC-ID), Martin Hecheltjen (ITM),
Barbara Kolany (ITM), Mykola Galushka (SAP), William Kilbride (DPC), Andi Rauber (SBA)

1 Executive Summary

This report outlines the huge societal benefits and commercial potential of being able to deliver a term digital preservation solution into the huge global marketplace and the challenges faced by digital preservation research; focusing on the developments which are of most significance over the last 12 months. This report was produced by the TIMBUS project (<http://timbusproject.net/>). TIMBUS is an EC funded project whose vision is to bring digital preservation into the realm of Business Continuity Management by developing the activities, processes and tools that ensure long-term continued access to business processes and the underlying software and hardware infrastructure.

The report highlights some interesting trends. Notable among these is the growing momentum behind the Internet-of-Things (IoT) economy. The wider diversity in consumer devices, coupled with the emerging IoT market suggests that we can expect to see shorter lifecycles for devices. This trend equally affects the lifespans of the technologies used by those devices.

The TIMBUS consortium continues to assert its position that the rapid pace of technology change significantly adds to severity and complexity of the digital preservation challenge making it more critical than ever before to address

Table of Contents

| | |
|--|-----------|
| 1 EXECUTIVE SUMMARY | 2 |
| TABLE OF CONTENTS | 3 |
| LIST OF FIGURES | 4 |
| LIST OF TABLES | 5 |
| LIST OF ACRONYMS | 6 |
| 2 INTRODUCTION | 7 |
| 3 THE DIGITAL UNIVERSE..... | 8 |
| 3.1 INTERNET USERS | 8 |
| 3.2 NEW DEVICES | 10 |
| 3.3 SMARTPHONE, TABLETS AND E-READER GROWTH..... | 12 |
| 3.4 WEARABLES | 15 |
| 3.5 THE SOCIAL NETWORK | 18 |
| 3.6 THE MOBILE NETWORK | 23 |
| 3.7 GROWTH OF EMBEDDED DEVICES/INTERNET-OF-THINGS | 27 |
| 3.8 CONCLUSIONS | 29 |
| 4 GLOBAL ARCHIVES | 30 |
| 4.1 WORLDWIDE ENTERPRISE STORAGE..... | 30 |
| 4.2 THE SOFTWARE LANDSCAPE | 31 |
| 4.3 TIERED/COLD STORAGE | 32 |
| 4.4 ARCHIVAL PROBLEMS FACING ORGANISATIONS | 34 |
| 4.5 USING CLOUD FOR ARCHIVING & DIGITAL PRESERVATION | 36 |
| 4.6 CONCLUSIONS | 39 |
| 5 INFLUENCE OF EMERGING TECHNOLOGIES | 40 |
| 5.1 INTEL DISTRIBUTION OF HADOOP, SAP HANA | 41 |
| 5.1.1 <i>Real-World Messaging Implementations</i> | 42 |
| 5.2 EMERGING BIG DATA TECHNOLOGIES | 43 |
| 5.2.1 <i>Cyc & OpenCyc</i> | 43 |
| 5.2.2 <i>COLIBRI, Apache Jena, OntoGen and OntoBridge</i> | 44 |
| 5.2.3 <i>Other Emerging Tools</i> | 46 |
| 5.3 CONCLUSIONS | 47 |
| 7 CONCLUSION | 48 |
| APPENDIX A: MARKET-WATCH BIBLIOGRAPHY (LINKS TO REFERENCED MATERIAL)..... | 49 |
| APPENDIX B: BIBLIOGRAPHY (LINKS TO REFERENCED MATERIAL) | 51 |

List of Figures

| | |
|--|----|
| Figure 1: World Internet Usage 2003-2013 | 8 |
| Figure 2: Percentage of Population using the internet 2005-2013 (source: http://www.itu.int/en/ITU-D/Statistics) | 9 |
| Figure 3: Internet User Distribution by time zone..... | 9 |
| Figure 4: Global IP Traffic 2011- 2016 | 10 |
| Figure 5: Technology Lifetime Cycles | 11 |
| Figure 6: Global PC and Tablet Shipments Q1 1995 to Q1 2013 | 12 |
| Figure 7: iPad vs iPhone Units Shipped Q1 1995 to Q1 2013..... | 13 |
| Figure 8: Device ownership (US market, 2013)..... | 13 |
| Figure 9: Updated Device Ownership Figures (source: PEW Research Center) | 14 |
| Figure 10: Print Books vs. E-books (source: Kayte Korwitts SurveyMonkey Blog) | 14 |
| Figure 11: Why do you own an E-reader | 15 |
| Figure 12: The big names in IT were all active in the new Wearables market in 2013 | 16 |
| Figure 13: KPCB Growth in Wearable Applications | 17 |
| Figure 14: Number of <i>MyFitnessPal</i> API Calls Oct 2012 Apr 2013 | 17 |
| Figure 15: Percentage who share 'everything' or 'most things' online..... | 18 |
| Figure 16: Selected Data from KPCB Report on Internet Trends 2013 | 20 |
| Figure 17: WW1 Soldier Diaries online..... | 21 |
| Figure 18: FCW Article on the Next Generation of Archiving..... | 22 |
| Figure 19: Architecture Overview of Memento Framework..... | 23 |
| Figure 20: Mobile Global Web Usage..... | 24 |
| Figure 21: Mobile Market Research (sources quoted in each, images produced by KPCB Internet Trends Report 2013)..... | 25 |
| Figure 22: Percentage of Global GDP 1820 to 2012 (image sourced from KPCB Internet Trends 2013 report) | 26 |
| Figure 23: Wintel, Mobile OS, Browser and Language usage over time | 27 |
| Figure 24: IoT; Proliferation across industry | 28 |
| Figure 25: Number of M2M connected devices and M2M traffic ²⁸ | 29 |
| Figure 26: IDC Storage Capacity Revenue & Capacity Shipped | 30 |
| Figure 27: Cost per GB from BackBlaze and IDC (inset graph is a composite of figures provided in table 9 on page 42 of that report for this 2012 and revised in 2013) | 31 |
| Figure 28: Storage Software Revenue | 32 |
| Figure 29: IDC Definition of 'Cold Storage' | 33 |
| Figure 30: An Enterprise Classification Scheme for Cold Storage ³⁷ | 33 |
| Figure 31: Characteristics of Media used in Archives ³⁷ (source: IDC) | 34 |
| Figure 32: Responses on dealing with Data Growth (source Informatica)..... | 35 |
| Figure 33: Responses on Archiving..... | 36 |
| (source: Informatica) | 36 |
| Figure 34: IBM Risk Global IT Risk survey 2010 | 36 |
| Figure 35: Wired.com article explaining the 3 V's of Big-Data..... | 40 |
| Figure 36: How Internet Search Engines Work (source: howstuffworks.com) | 41 |
| Figure 37: The Intel-HANA Story | 42 |
| Figure 38: IDH Architecture for Intel Manager..... | 42 |
| Figure 39: Other State-of-The-Art (SoTA) Implementations ⁵⁰ | 43 |
| Figure 40: Cycorp and Cycorp Europe Homepages | 44 |
| Figure 41: Jena, OntoGen, OntoBridge and COLIBRI..... | 45 |
| Figure 42: Output of Enrycher Web API | 47 |

List of Tables

| | |
|---|----|
| Table 1: Top 15 Social Media Sites (Feb 2014) | 19 |
| Table 2: Top 5 questions to ask before outsourcing archival to a cloud vendor | 37 |

List of Acronyms

| | |
|----------|---|
| BBC | British Broadcasting Company |
| BCM | Business Continuity Management |
| BCP | Business Continuity Planning |
| CAD | Computer Aided Design |
| CAGR | Compound Annual Growth Rate |
| CMS | Caixa Magica Software |
| CSL | Cloud Services Lab |
| DOT | Development Opportunity Tool |
| EAB | External Advisory Board |
| EIT | European Institute of Innovation & Technology Foundation's |
| EPS | Emergency Planning Solutions. A consultancy and training firm in Belfast |
| IT-EC | IT Engineering Computing. An internal IT support division within Intel Corporation. |
| EMC | EMC Corporation (E=MC ²) |
| ESG | Enterprise Storage Group – A provider of industry analysis data |
| ESL | Energy and Sustainability Lab |
| GB | Gigabyte |
| GWAP | Game With A Purpose |
| IDC | International Data Corporation -- A provider of industry analysis data |
| iERM | Intelligent Enterprise Risk Management |
| IoT | Internet-of-Things |
| IP | Intellectual Property |
| IT | Information Technology |
| iLE | Intel Labs Europe |
| ILM | Information Lifecycle Management |
| JPF | Joint Path Finding |
| KPCB | Kleiner Perkins Caufield & Byers |
| OAIS | Open Archival Information System |
| OEM | Original Equipment Manufacturer |
| PASIG | Preservation and Archival Special Interest Group |
| PB | Petabyte |
| PCC | Project Coordination Committee |
| R2 | Risk and Resilience Ltd. An organisation based in Belfast |
| INESC-ID | INSTITUTO DE ENGENHARIA DE SISTEMAS E COMPUTADORES, INVESTIGACAO E DESENVOLVIMENTO |
| Intel | Intel Performance Learning Solutions Limited |
| ITM | Institute for Information-, Telecommunication- and Media Law Muenster |
| KIT | Karlsruher Institut für Technologie |
| LTDP | Long Term Digital Preservation |
| MIP | Mega-Impact Project |
| SAP | SAP AG |
| SBA | Secure Business Austria |
| SDI | Software Defined Infrastructure |
| SI | System Integrator |
| SLA | Service Level Agreement |
| SME | Small Medium Enterprise |
| SNIA | Storage Network Industry Association |
| SQS | Software Quality Systems |
| STEP | Standard for the Exchange of Product Model Data |
| TCO | Total Cost of Ownership |
| VNI | Cisco's Visual Networking Index |
| WP | Work Package |

2 Introduction

Since the late 1990's, and thanks in part to Moore's law, technological progress has accelerated, resulting in lower product costs, mass affordability and widespread adoption of digital technology. Global competition and open standards have resulted in product innovation, shorter product lifespans and in increased rates of technological obsolescence which in turn have raised the urgency of addressing the fundamental problems of digital preservation.

The market assessment carried out in this report, checks the current state of play as 2014 draws to a close. There seems to be little sign of this trend slowing, in fact the opposite is the case. All predictions today are of an even more highly technologically integrated society than at any earlier point in history but this is happening through an unprecedented diversity of devices which is challenging the concept of how we think of traditional computing. The exact shape of the future is constantly changing, but never before has there been such potential for market disruption as there is today.

Against that ever accelerating adoption of technology, the European Commission continues to recognise the digital preservation challenge and through funding projects such as TIMBUS it is planting the seeds that will grow to address this multifaceted and dynamic need.

The information presented in this report was compiled from primary sources by the EC co-funded TIMBUS FP7 project to inform the reader. Those sources are quoted in every instance. The TIMBUS project seeks to bring digital preservation more into mainstream applying concepts of business continuity and risk management to the domain of digital preservation.

The goals of research initiatives such as TIMBUS remain ambitious in that they face many challenges in absorbing, integrating and building upon the considerable body of work already completed by numerous previous projects and agencies in many aspects of Digital Preservation, and even beyond that in related disciplines of big-data technologies and emerging meta-data generation techniques. This report should help any reader interested in the area by pulling together many such sources of information into one place and adding an assessment of these from the perspective of the TIMBUS FP7 project.

3 The Digital Universe

A convincing body of evidence exists to relate the growing tsunami of data both in terms of number of new users, new devices, usage models, all contributing to the vast amounts of data being generated, transmitted, processed, presented and stored. Data growth is a large driver behind the need for digital preservation. Some of the main reasons for this data growth are the increasing number of people using the internet worldwide, growth in number of devices generating data, richer data being generated and by new usage models served by new devices.

This chapter will consider the predictions and trends of 12 months ago and report on how they have changed. Additionally, it will take into consideration how the penetration of internet connectivity is changing our daily habits. This is interesting because it is not only the growth in internet usage that drives the requirement for long-term preservation but also the changes it has enabled in how we go about our everyday lives.

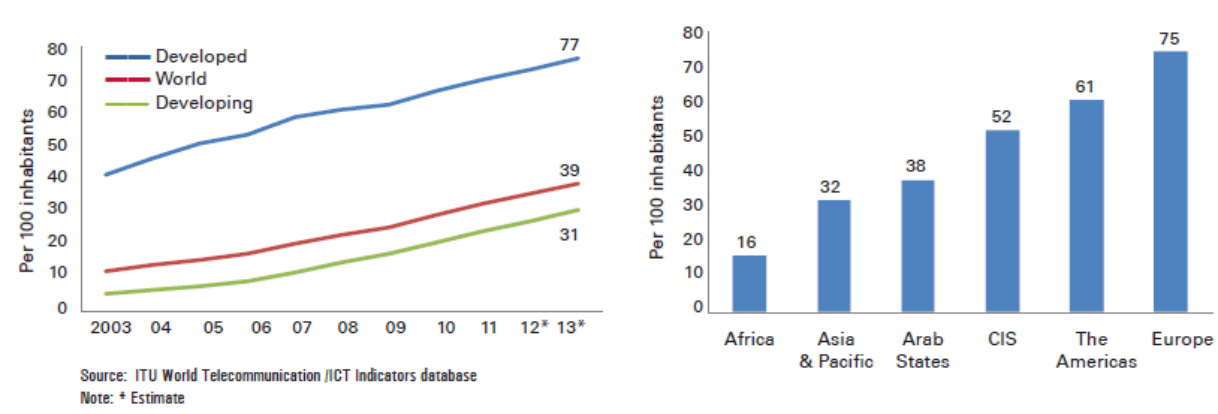
The reason this sort of data is important is because it paints a picture of how our lives are changing as technology becomes more persuasive. The statistics below show that as we spend more of our time online, we are also generating huge volumes of data. The more data we produce, then the more data we potentially need to preserve.

3.1 Internet Users

There are many sources, among them the World Internet Statistics¹ group whom have stated figures such as “it is estimated that there will be 2.2 billion Internet users in the world” by 2013. As shown below in Figure 1, according to the ITU Telecommunications Development Bureau¹, there are 2.7 billion, or almost 40% of the world’s population online in 2013. In 2011, the figure was 28% as shown in Figure 2.

2.7 BILLION PEOPLE – ALMOST 40% OF THE WORLD’S POPULATION – ARE ONLINE

Internet users by development level, 2003-2013*, and by region, 2013*



In 2013, over 2.7 billion people are using the Internet, which corresponds to 39% of the world’s population.

In the developing world, 31% of the population is online, compared with 77% in the developed world.

Europe is the region with the highest Internet penetration rate in the world (75%), followed by the Americas (61%).

In Africa, 16% of people are using the Internet – only half the penetration rate of Asia and the Pacific.

Figure 1: World Internet Usage 2003-2013¹

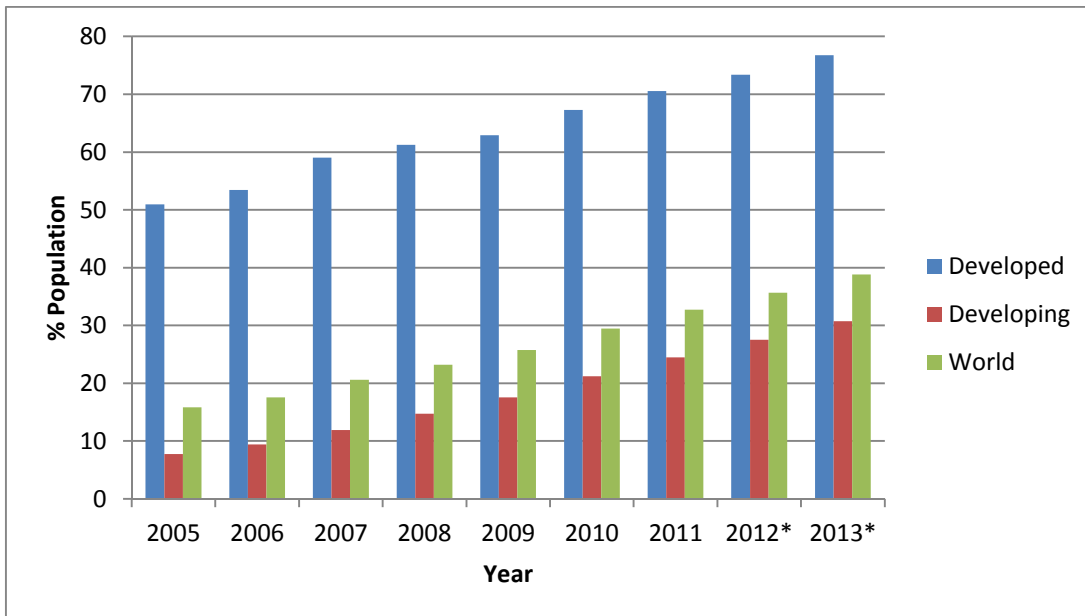


Figure 2: Percentage of Population using the internet 2005-2013

(source: <http://www.itu.int/en/ITU-D/Statistics>)

Asia and Africa continue to be the largest growth areas as internet penetration in those regions lags behind the world average at 32% and 16% respectively. But, as shown in Figure 3, these regions are well represented globally if viewed in terms of raw numbers of internet users instead of percentages of the population online.

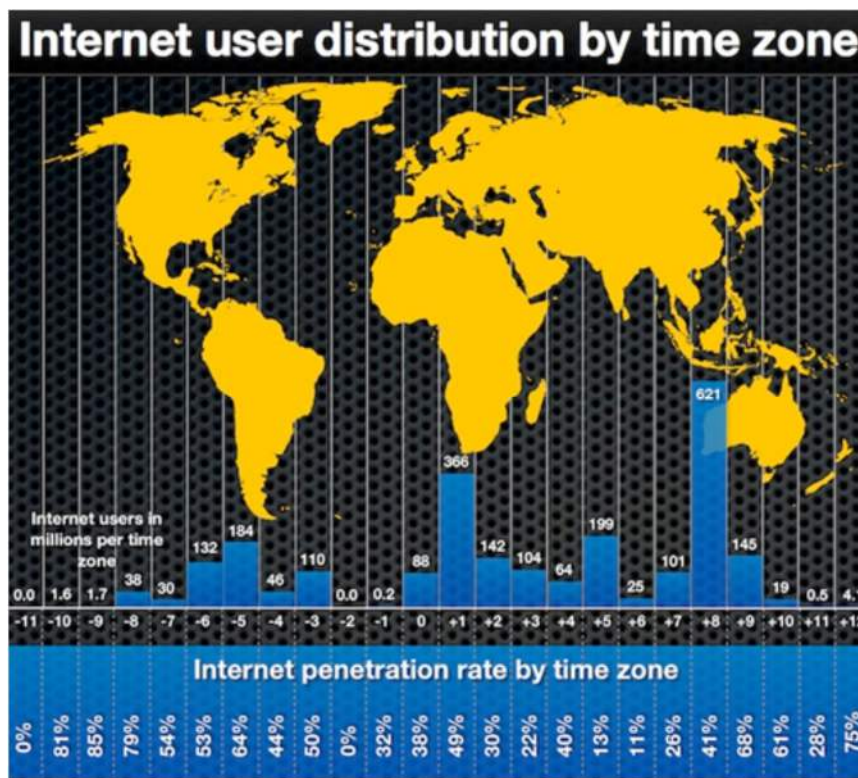


Figure 3: Internet User Distribution by time zone²

With the increase in users comes a direct increase in network traffic as shown below in Figure 4. In year 1 of the TIMBUS project, this deliverable produced a graph of global internet traffic from Cisco's VNI index showing traffic rising from about 2,500 petabytes per month in 2005 to just over 20,000 petabytes (or 20 Exabyte's) in 2010. Last year, the graph of global internet traffic from Cisco's VNI index on the left shows

how the traffic has been rising, with traffic in 2011 just under 30 Exabyte's and growing to over 107 by 2016. The latest figures available from Cisco in 2014 show the trend continuing with the largest data types being video and video communications with web traffic and file sharing also being significant contributors. The data being produced shows that in just five short years, from 2012 through to 2017, internet traffic will have increased 3-fold. This rapid pace of growth is even affecting the ability of Telco's to provision capacity to meet this demand.



Figure 4: Global IP Traffic 2011- 2016³

To put this incredible growth in perspective, world internet penetration at the end of 2013 is still only about 38%. This shows that there is still huge potential for growth in terms of internet users which in turn will drive new usage models. As humans become increasingly internet connected, it is reasonable to conclude that there will be increasing difficulties in identifying the subset of personal and business data that needs to be preserved and carrying out that preservation in a cost effective and efficient manner. The next sections of the report will examine what has changed in 2013, in how we are interacting with the internet to give us an understanding of why the need to preserve data for the long term may be increasing.

3.2 New Devices

New devices present a particular challenge for digital preservation. In fact, they create a moving target of increasingly diverse, more mobile hardware platforms whose life spans are shorter than traditional IT infrastructure and whose operational profiles are extremely difficult to predict in the medium term and almost impossible if projected out over a number of years. A noticeable new trend in 2013 is how the established paradigm of technology cycles lasting for about 10 years, as shown below in Figure 5 from a KPCB report⁴ on Internet Trends in 2013 is coming under threat by new devices.

Technology Cycles Have Tended to Last Ten Years

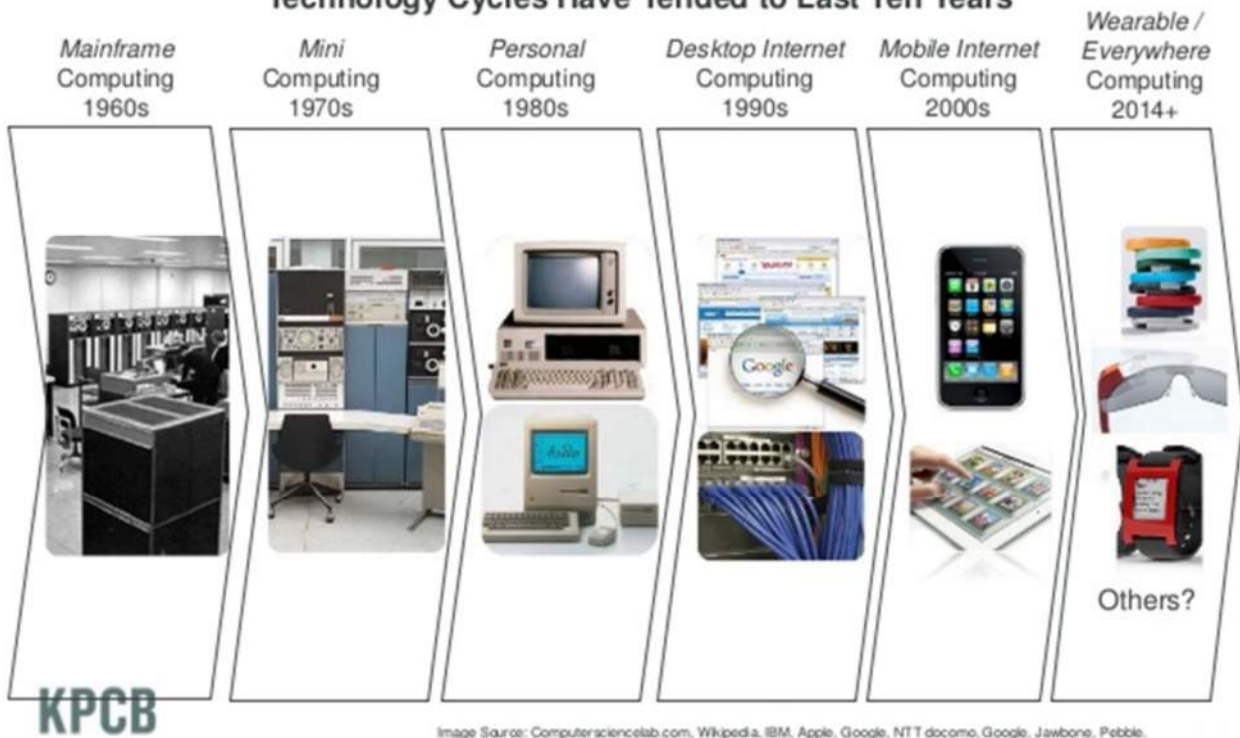


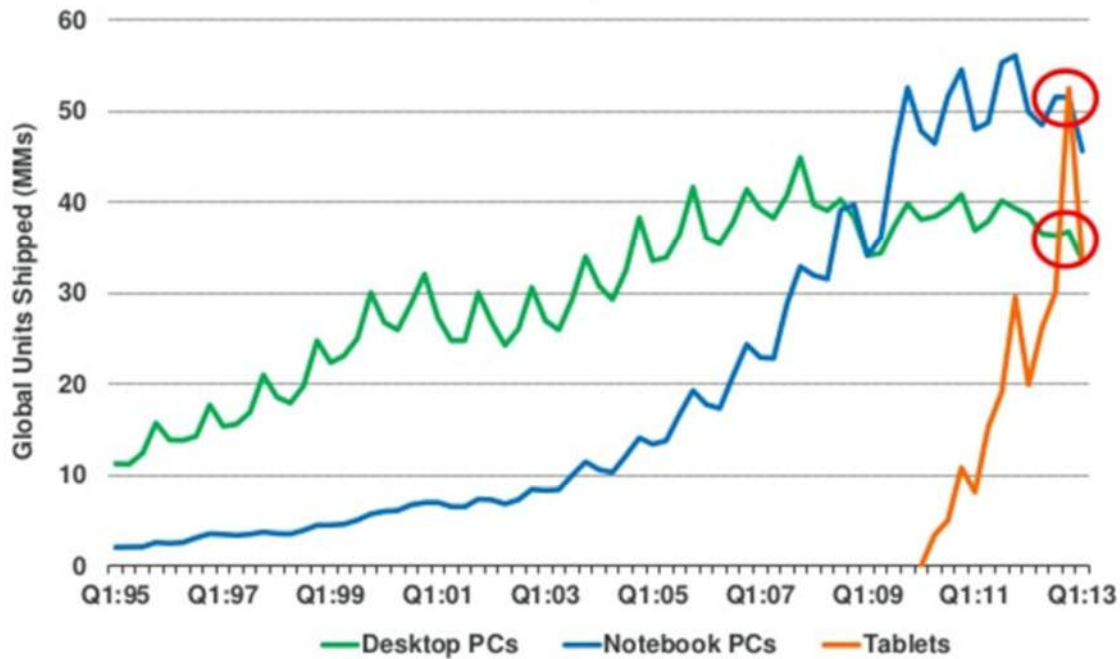
Image Source: Computersciencetab.com, Wikipedia, IBM, Apple, Google, NTT docomo, Google, Jawbone, Pebble.

Figure 5: Technology Lifetime Cycles

(source: KPCB Report on Internet Trends 2013)

In the past, these major 10 year cycles have coincided with bursts of new technology users who adopt new devices and bring them in to their everyday lives. This is because these devices offer compelling new usage models which are enabled by new functionality, lower price points and new form factors.

In the same KPCB report, Figure 6 is presented which shows the speed at which new device technologies have ramped in recent years. We can see that tablet shipments have surpassed PC and laptop shipments in only three years. This trend suggests that the newer markets are ramping much more rapidly than the traditional ones. In an interview on April 29th 2013, former BlackBerry CEO, Thorsten Heins said *"in five years I don't think there'll be a reason to have a tablet anymore⁵"*. In BlackBerry's case, they are questioning if tablets are a viable long-term business model. While it remains to be seen if these sorts of considerations will result in even shorter technology lifecycles, what is certain is no one in the IT sector is taking it for granted that any particular technology or form factor will be around in more than a few years' time. Each form factor and each device is facing an extremely competitive future and ultimately it is technology consumers, and not the CEO's who will decide. A major impact out of this trend for research initiatives such as TIMBUS is that the shorter these technology trends become, then the more difficult the long-term preservation challenge will become.



KPCB

Note: Notebook PCs include Netbooks.
 Source: Katy Huberty, Ehud Gebilum, Morgan Stanley Research, Gartner. Data as of 4/13.

Figure 6: Global PC and Tablet Shipments Q1 1995 to Q1 2013

3.3 Smartphone, Tablets and e-Reader Growth

Smartphones, tablets and e-readers are another subset of new devices which deserve some special focus. Figure 7 below, shows figures published by Apple for global shipments of iPads and iPhones for the 12 quarters after the initial product launch. The figures back up what we have seen previously in section 3.2 above, namely that the adoption of new devices is happening more rapidly with each new technology trend. In the same way that Figure 6 showed how rapidly tablets have caught up with PC and laptop unit sales, tablets, have also ramped more quickly than their direct predecessor, the smartphone.

A rapid adoption of notebook PC's from 2008 onwards had a direct impact on finally halting the growth of the PC market but when tablets enter the market in 2010, the combined effect on the traditional PC is truly disruptive, at least it would appear to be so at this early stage in the product lifecycles. It remains to be seen how resilient the PC and Notebook markets will be to these new challengers and if they will disappear entirely or merely level off at some steady state beyond which their unit sales will not fall too far.

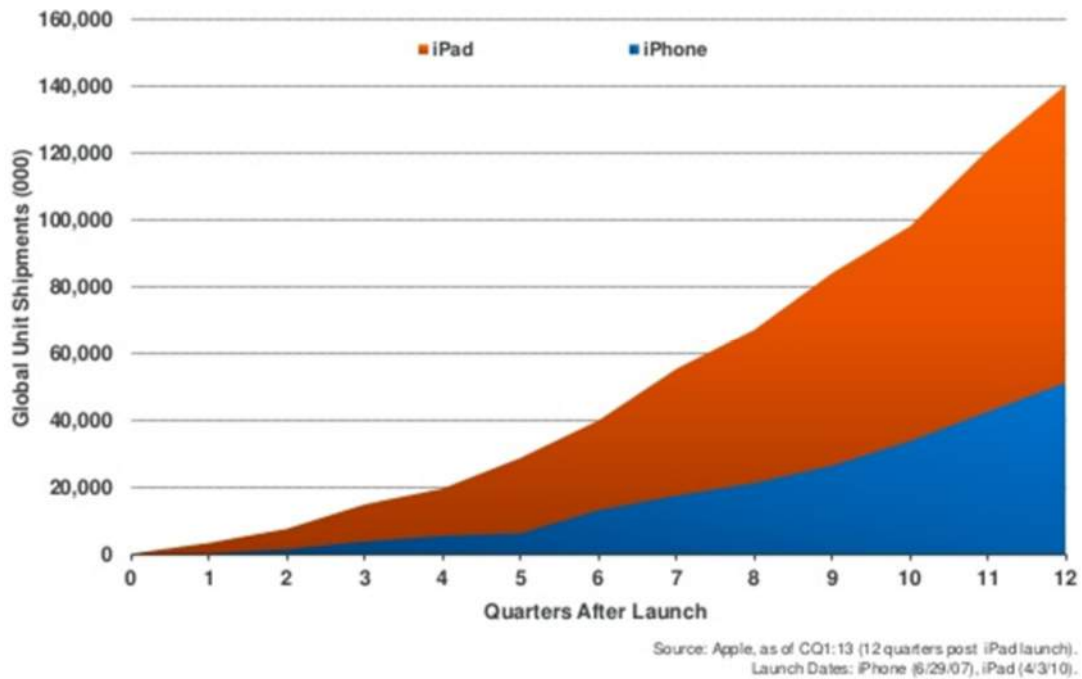


Figure 7: iPad vs iPhone Units Shipped Q1 1995 to Q1 2013

Figure 8 below shows the device ownership figures which were available for American adults (18 years+). As 2014 passes by, updated figures are available and are shown in Figure 9.

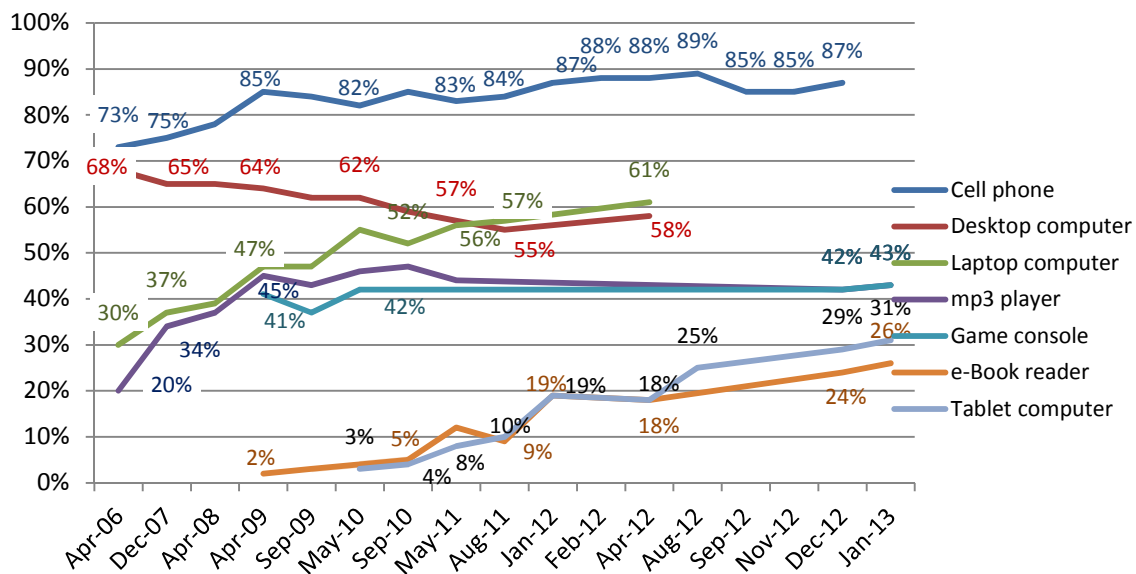


Figure 8: Device ownership (US market, 2013)⁶

As can be seen in the new figures, newer devices are seeing growth. If we examine the figures more closely, it could probably be safely assumed that cell phones have a saturated market. Desktop PC's and smartphones are starting to lose out to laptops and tablets although the figures don't make it clear in what precise proportions those may be happening and eBook readers are continuing to rise in popularity, probably at the cost of traditional media such as books. Even the adoption of eBooks is not something to be ignored because it suggests that behind the scenes there is growing momentum behind the migration from a relatively long-lived media such as paper to a short lived one in electronic formats. These sorts of social trends are important considerations if we are to preserve a representation of old literature on future devices. This is a problem which memory institutions have been tackling for hundreds for years as they preserve writings that date back to the middle ages and beyond.

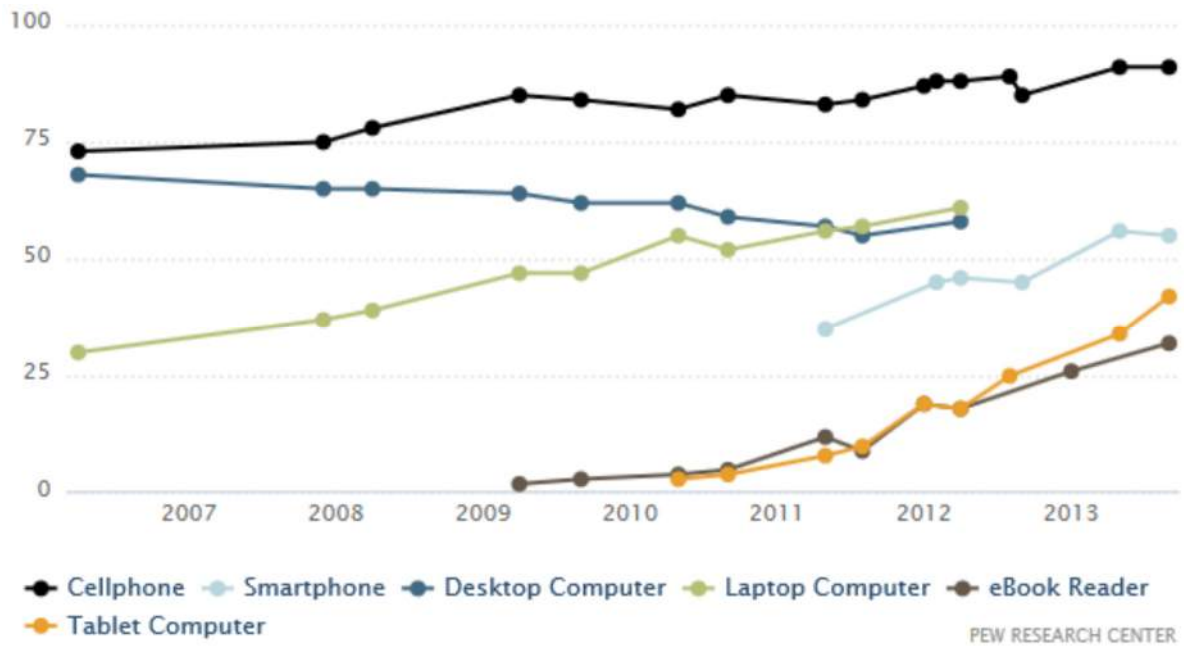


Figure 9: Updated Device Ownership Figures
(source: PEW Research Center)

There are many sources available discussing the future of reading at present and trying to understand the effect of the apparent migration from books to E-books. A recent *SurveyMonkey*[®] blog⁷ by Kayte Korwitts contributes to that discussion and some of their findings are shown below in Figure 10. *SurveyMonkey* asked 300 American readers about their habits and views on this trend.

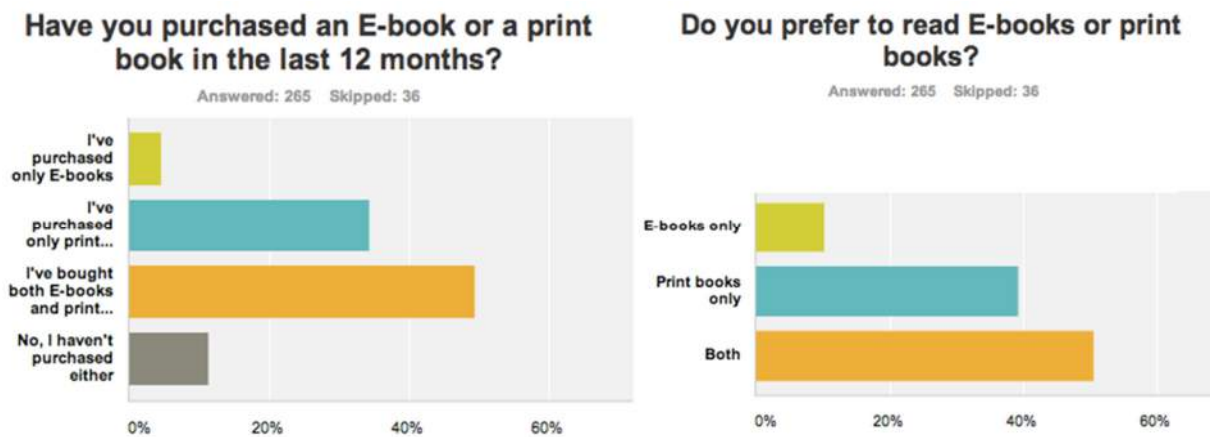


Figure 10: Print Books vs. E-books
(source: Kayte Korwitts SurveyMonkey Blog)

The results showed that there is a large trend taking place but it's not necessarily the one that one may think at first glance. Nor is it having a negative effect on either the number of people reading. In fact, at present, if the views are widely held by readers, the trends do not look like they will rapidly replace paper as a media. When this was explored further, the largest reasons for owning an E-book were for convenience, often when commuting as shown below in Figure 11. The inability in the case of E-books to entirely push paper-based books to the side is that people still prefer the experience of turning pages and the touch and feel of a book rather than just having the material available at their fingertips.

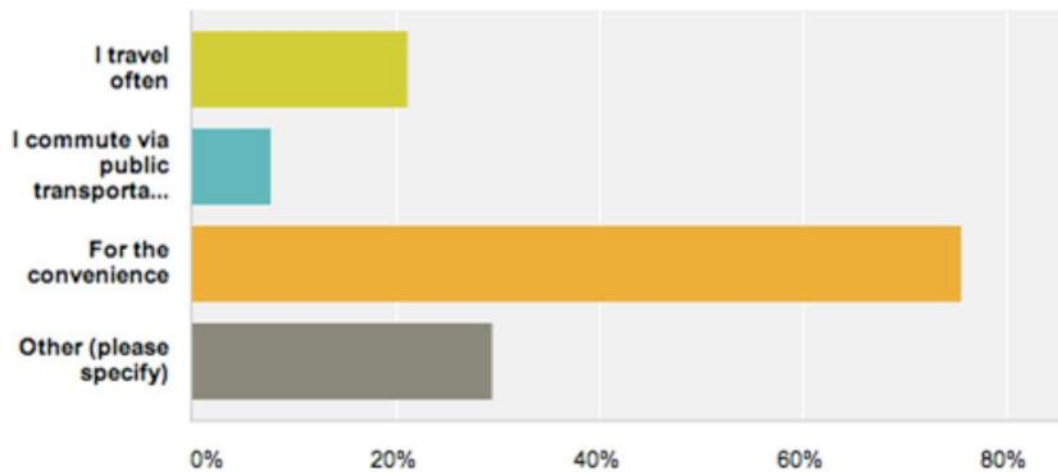


Figure 11: Why do you own an E-reader

This raises an interesting consideration from a Digital Preservation research perspective. When we consider a business process running on a certain set of infrastructure or devices today, do we adequately consider the user experience when attempting to preserve and re-create the business process in the future? As can be seen from the example of books, it is probably fair, and even prudent, to not discount this aspect of our interactions with technology. Going further, when we consider how rapidly new devices such as mp3 players, smartphones and tablets have gained widespread adoption, we must accept that this is also largely due to the desirable form factor and the user experience from interacting with these devices in their native forms.

3.4 Wearables

Wearables have been around for a while, but this market has been particularly topical in 2013 with large players such as Intel, ARM, Sony, Texas Instruments and others all launching enabling products during the year, some of which are shown below in Figure 12. ‘Wearables’ is the term used to refer to an emerging market where compute devices become more personal than ever before. It is a sector that is being enabled by increasingly smaller compute devices such as the Intel® Galileo® and Intel® Edison® SoC’s (System on Chip) which are powered by tiny, but powerful, microprocessors such as the Intel® Quark®.

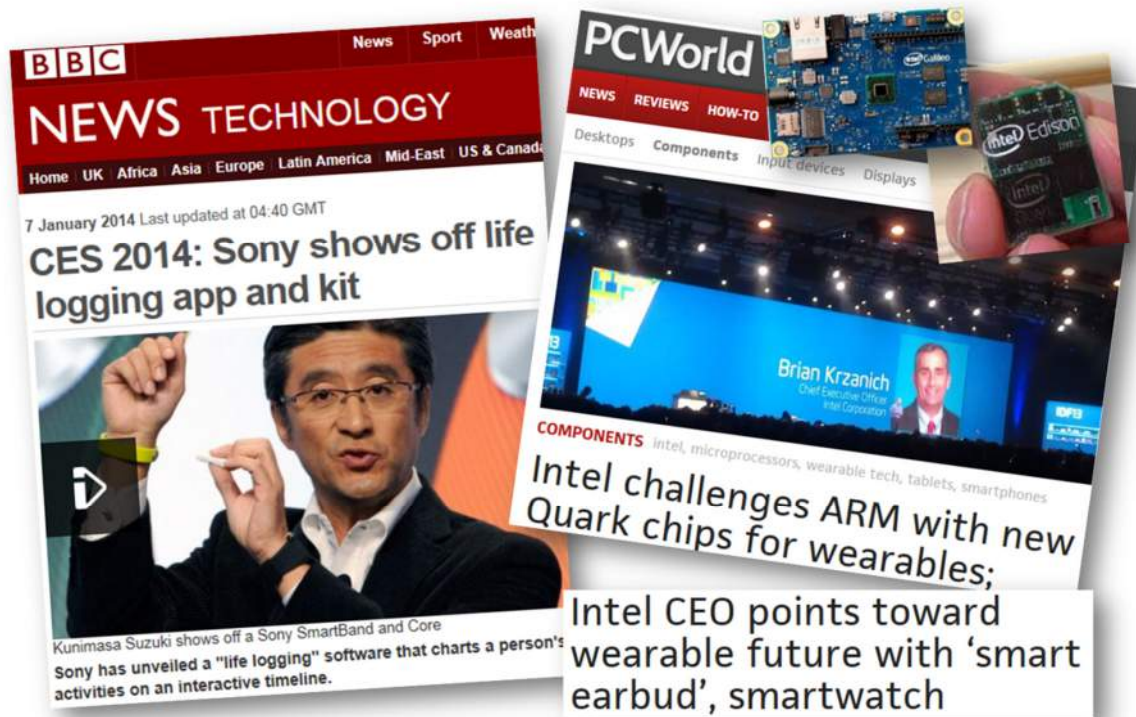


Figure 12: The big names in IT were all active in the new Wearables market in 2013

All of the big names in the IT industry are looking at this emerging market and trying to influence its direction by launching devices, applications and experiences which they hope will be disruptive game-changers.

In the Digital Preservation research world, these devices are presenting yet another new challenge. These devices are interesting because they are contributing to acceleration in the pace of technology, which in turn are exacerbating the long-term digital preservation challenge. These new devices cannot be considered 'dumb' consumers of data, but, to take one example, in Life-logging, they are actually producing and potentially processing large volumes of data to be uploaded to the cloud for some future use. Other examples shown below in the KPCB report on Internet Trends 2013 in Figure 13 point to the growing number of applications and uses that this data is being put to.

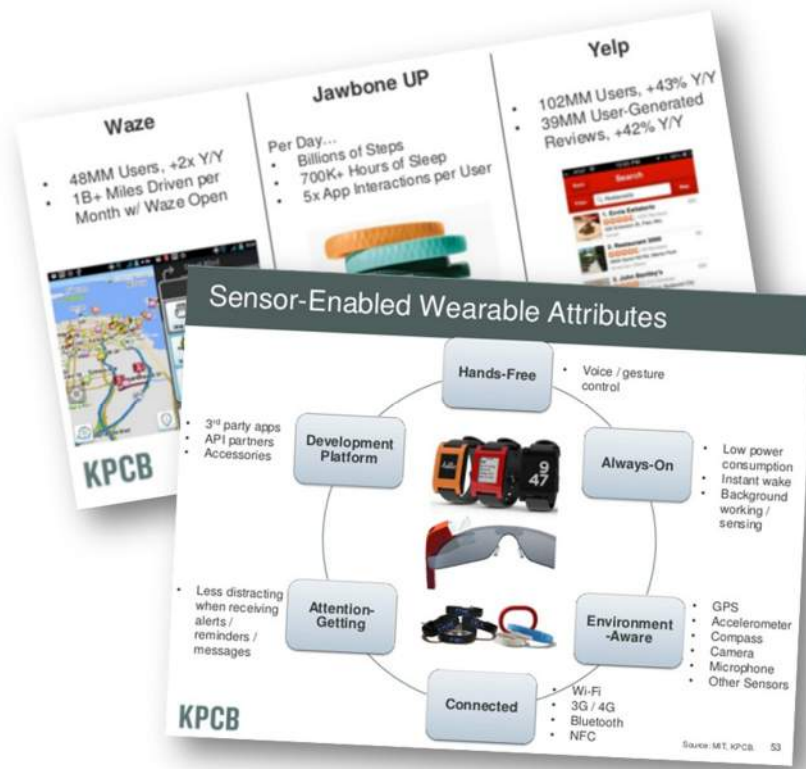


Figure 13: KPCB Growth in Wearable Applications

Another good example of the growing adoption of wearables is shown in the same KPCB report using the example of the *MyFitnessPal*[®] application, shown below in Figure 14. The number of API calls made to the application is an indication how quickly these new use models can be adopted.

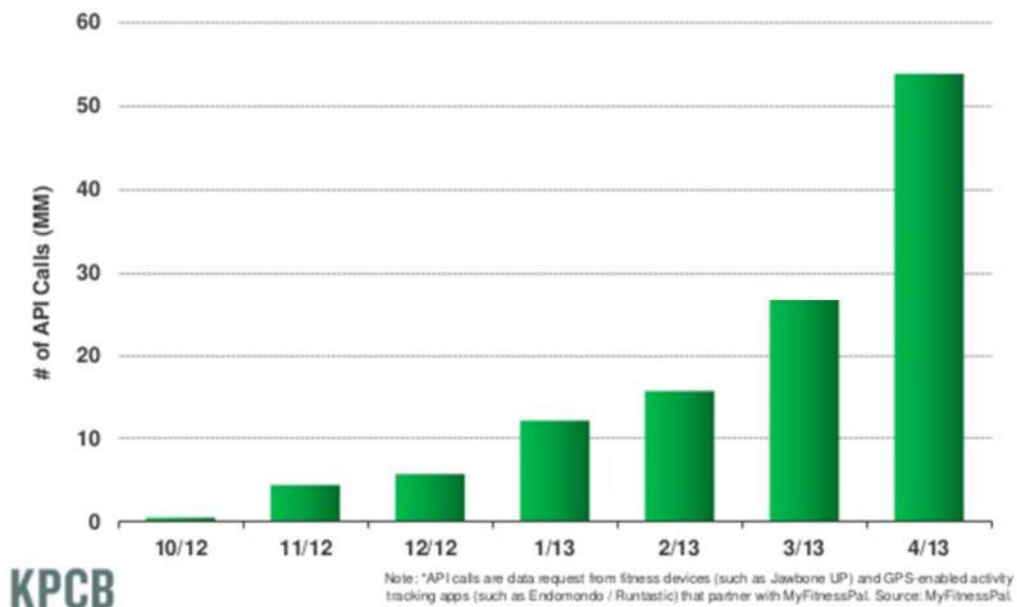


Figure 14: Number of MyFitnessPal API Calls Oct 2012 Apr 2013

An important question is whether the data that lives in any of these applications has long term value. On one hand, data about an individual such as captured by a wearable which feeds an application such as *MyFitnessPal* might seem trivial and of little value to anyone other than the individual concerned and even then, only at a certain point in time. The reality is that this data is actually highly valuable to a lot of

people and not just in the short term, but over as long a period of time, and from as geographically dispersed an area as possible. There are many health research projects funded globally that use small study groups to try to ascertain trends about, to pick just one example, obesity. The type of data we are talking about from Wearables, if it were shared in some way and if it was sufficiently anonymised to take in to consideration any privacy concerns, is actually highly valuable research data which would be either impossible, or cost prohibitive to otherwise gather in any single study.

Chris Woods of Intel Labs Europe spoke on the topic of *Data as a Service for Social Change*⁹ at the European Institute of Innovation & Technology Foundation’s (EIT) forum on Innovation. His full talk and materials are available on the EIT website. In a very compelling keynote, Chris explains exactly how this data could be made available in a way which does not actually expose the privacy of the data but yet still would allow researchers to have access to vast datasets which they could never hope to otherwise gather through the approach that is used today with small scale funding of limited population pools, all done in typically geographical proximity to one another.

In the context of Wearables, there are similar considerations about the data generated, its value as a research tool, how that data can be preserved and what software environments are required to interpret and render that data. Needless to say, all this must be undertaken in a legally approved way. However, legislation at present and by its very nature, struggles to keep pace with a rapidly evolving technology landscape. Despite the varying degrees of legal protection offered in different jurisdictions, or perhaps because of them, Figure 15, below, indicates that individual preferences for sharing data about themselves online varies significantly across geographical regions today. But it does show that people are willing to share at least some data under whatever conditions they had undertaken to do so. This data doesn’t necessarily have to be used for targeted marketing by large corporations but it has value at a social and health level as previously explained by Chris Woods of Intel.

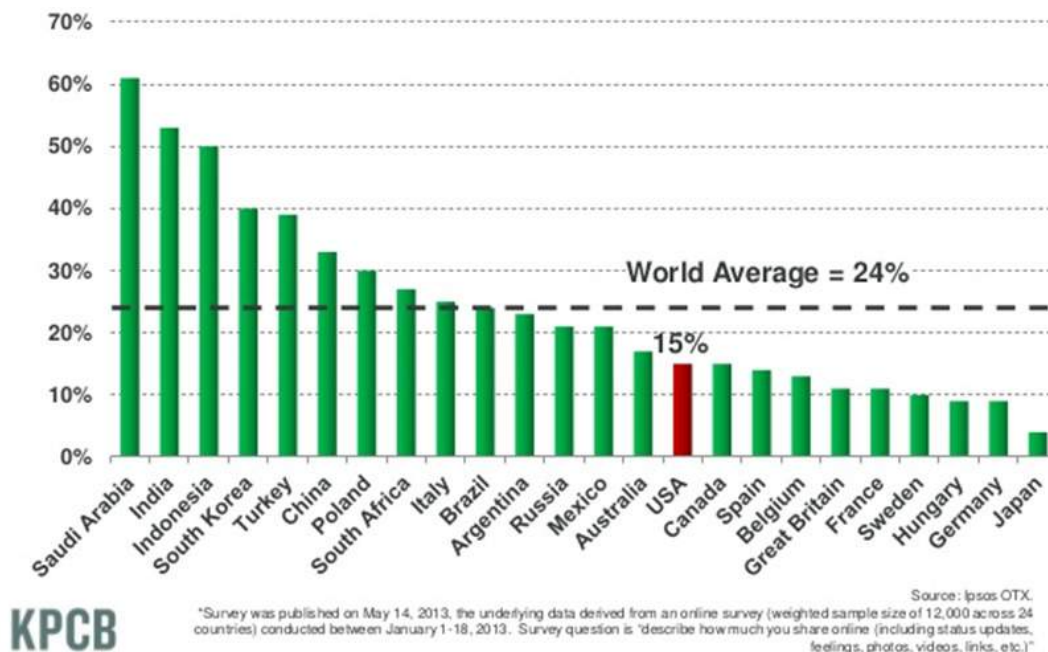











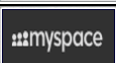





Figure 15: Percentage who share ‘everything’ or ‘most things’ online

3.5 The Social Network

Table 1 below shows the top 15 Social Media sites as of February 2014 ranked in order of estimated unique monthly visitors. This data is published by *The eBusiness Knowledgebase* on their website at ebizmba.com¹⁰. Data such as this is important to consider because it shows how our relationship with the internet is evolving over time. This is important from a Digital Preservation research perspective in order

to understand what sort of data we are generating and what combinations of software and hardware environments are being used to generate and consume that data.

Table 1: Top 15 Social Media Sites (Feb 2014)

| Ranking (Feb 2014) | Site | Stats |
|-----------------------|---|--|
| 1 |  | 3 - eBizMBA Rank 900,000,000 - Estimated Unique Monthly Visitors 3 - Compete Rank 3 - Quantcast Rank 2 - Alexa Rank. |
| 2 |  | 15 - eBizMBA Rank 290,000,000 - Estimated Unique Monthly Visitors 30 - Compete Rank 5 - Quantcast Rank 9 - Alexa Rank. |
| 3 |  | 16 - eBizMBA Rank 250,000,000 - Estimated Unique Monthly Visitors 24 - Compete Rank 17 - Quantcast Rank 8 - Alexa Rank. |
| 4 |  | 27 - eBizMBA Rank 150,000,000 - Estimated Unique Monthly Visitors 40 - Compete Rank 14 - Quantcast Rank 26 - Alexa Rank. |
| 5 |  | 30 - eBizMBA Rank 126,000,000 - Estimated Unique Monthly Visitors *32* - Compete Rank *28* - Quantcast Rank NA - Alexa Rank. |
| 6 |  | 31 - eBizMBA Rank 125,000,000 - Estimated Unique Monthly Visitors 55 - Compete Rank 13 - Quantcast Rank 25 - Alexa Rank. |
| 7 |  | 70 - eBizMBA Rank 100,000,000 - Estimated Unique Monthly Visitors 52 - Compete Rank 118 - Quantcast Rank 41 - Alexa Rank. |
| 8 |  | 94 - eBizMBA Rank 80,000,000 - Estimated Unique Monthly Visitors 115 - Compete Rank 102 - Quantcast Rank 65 - Alexa Rank. |
| 9 |  | 97 - eBizMBA Rank 79,000,000 - Estimated Unique Monthly Visitors *150* - Compete Rank *120* - Quantcast Rank 22 - Alexa Rank. |
| 10 |  | 284 - eBizMBA Rank 40,000,000 - Estimated Unique Monthly Visitors 22 - Compete Rank 70 - Quantcast Rank 761 - Alexa Rank. |
| 11 |  | 486 - eBizMBA Rank 38,000,000 - Estimated Unique Monthly Visitors 703 - Compete Rank 376 - Quantcast Rank 378 - Alexa Rank. |
| 12 |  | 518 - eBizMBA Rank 35,000,000 - Estimated Unique Monthly Visitors 687 - Compete Rank 546 - Quantcast Rank 321 - Alexa Rank. |
| 13 |  | 599 - eBizMBA Rank 34,000,000 - Estimated Unique Monthly Visitors 1,538 - Compete Rank 103 - Quantcast Rank 155 - Alexa Rank. |
| 14 |  | 1,148 - eBizMBA Rank 10,500,000 - Estimated Unique Monthly Visitors 1,325 - Compete Rank 177 - Quantcast Rank 1,941 - Alexa Rank. |
| 15 |  | 1,189 - eBizMBA Rank 10,000,000 - Estimated Unique Monthly Visitors 141 - Compete Rank 207 - Quantcast Rank 3,220 - Alexa Rank. |

It is also important to think about the effect on the volume and types of data being produced by these new usage models, backed up by Figure 16. YouTube continues to grow, Twitter has added photo and video support. Dropcam users now upload even more video content than YouTube while social media services such as Flickr, Snapchat, Instagram and Facebook are not only as popular as ever, but in terms of photo uploads, are actually still accelerating with no levelling off of demand in sight yet.

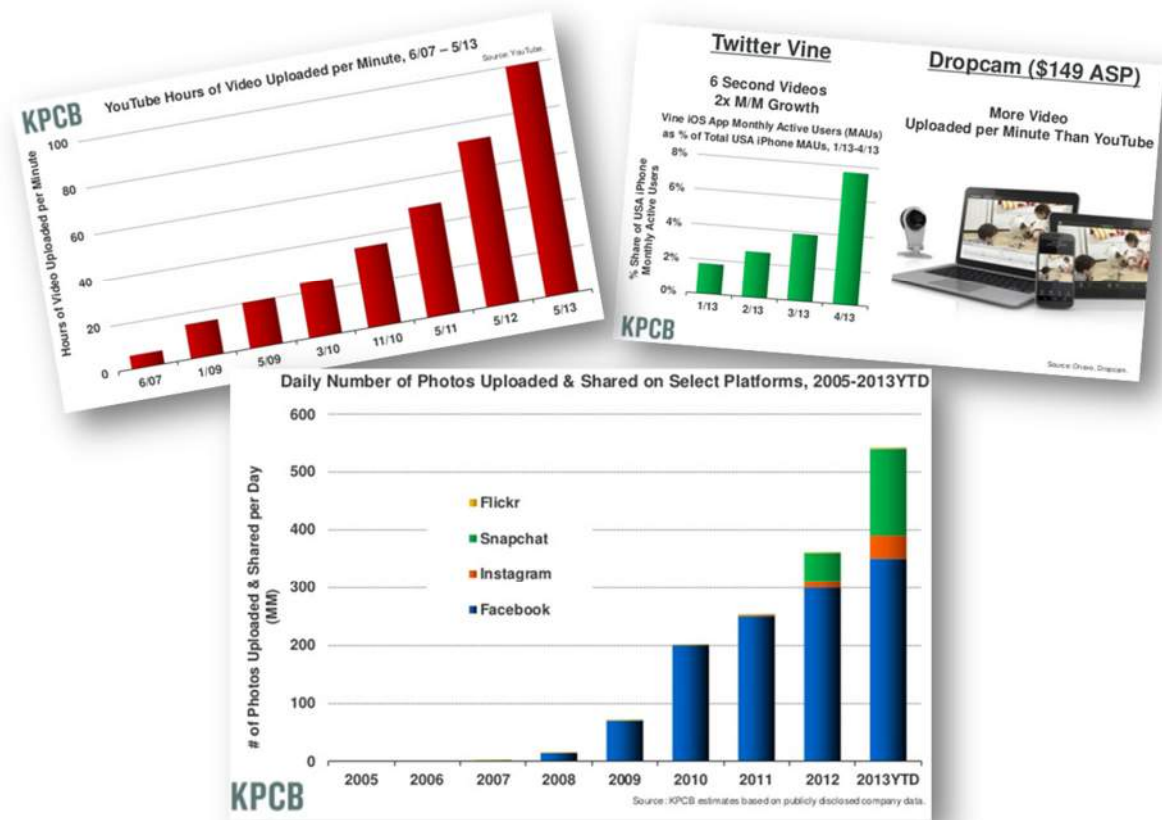


Figure 16: Selected Data from KPCB Report on Internet Trends 2013

At first glance, these services may seem unrelated to the area of Digital Preservation research. After all, the long-term preservation of static video and photos is a large part of the activities that have been traditionally undertaken by memory institutions and therefore should not be too challenging a problem to deal with. The reality, as is often the case, is not quite so straight forward.

Figure 17, below shows a recent article from the BBC which speaks about the digitising of diaries from soldiers who took part in World War 1. These diaries have been put online by the National Archives in the UK.

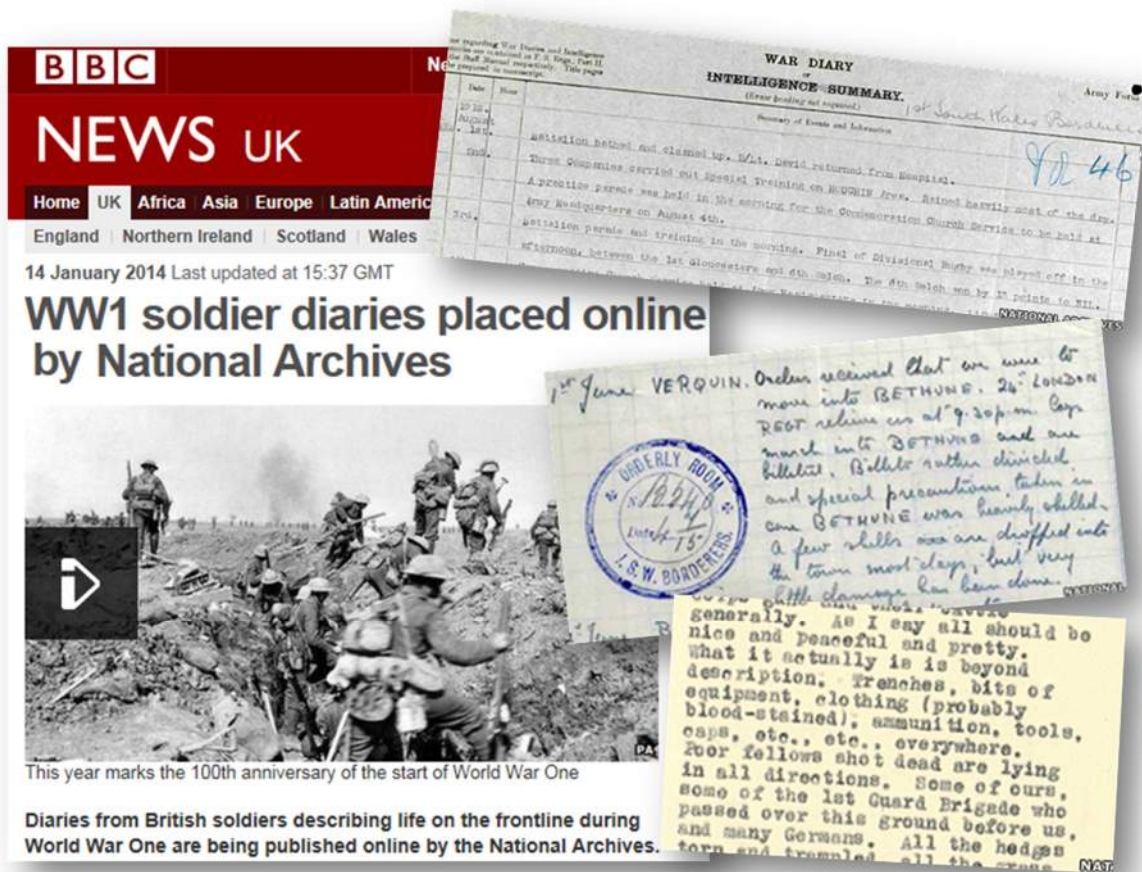


Figure 17: WW1 Soldier Diaries online¹¹

The National Archives are also taking a crowd-sourcing approach to digitising these records because they are of cultural interest. These diaries are invaluable accounts from individuals which are untainted by the propaganda that can surround war and not yet skewed through the eye of a revisionist who had the benefit of knowing how the historical events played out. As such, they are highly valuable. If we roll the clock forward to our present time in history and ask ourselves what is the modern day equivalent of such records, we are left with a potentially much different view of social media sites. Rather than just online playthings, social media sites are in fact capturing a first-hand image of ourselves at a certain point in time. It is not surprising to find an increasing number of articles such as the one shown below in Figure 18 about this topic. Richard Stapleton, senior deputy at the Digital Communications Division and senior Web strategist at the Department of Health and Human Services in that article is quoted as saying: *"This past year, our office began using Vine, Pinterest, Tumblr and Storify. Each of these platforms requires a new evaluation and approach."* Therefore, it would seem that the diversity and popularity of social media sites are making them increasingly relevant from a Digital Preservation research perspective. In the United States, the Federal Records Act¹² of 1950 established a framework for the administration of records in federal agencies. As government bodies around the world increasingly use social media outlets to perform certain public functions, the interpretation of what constitutes a federal record is broad. In the case of the United States, the 1950 Records Act says a record is any material "regardless of physical form or characteristics". This interpretation would easily extend to social media interactions such as tweets or blogs and puts an onus on federal agencies in the US to identify what constitutes the official record of their business and how it will be managed, up to and including any social media presence which an agency may have online. An IDC report¹³, titled *Worldwide Storage Software 2013–2017 Forecast and 2012 Vendor Shares* from October 2013 backed this up stating: *"Although in its infancy, growing litigation will drive significant increases in social media archiving within heavily regulated industries. Fueled by enforcement of regulations, organizations will increasingly store, monitor, and govern social media communications."*

Social media: The next generation of archiving

By John Moore Nov 25, 2013

Social media platforms such as Twitter, Facebook and YouTube have become commonplace tools for government outreach. Agencies

to "The social media landscape changes frequently, and the tools and platforms that we use to di-engage with the public constantly evolve," said Richard Stapleton, senior deputy at the Digital Communications Division and senior Web strategist at the Department of Health and Human Services. "This past year, our office began using Vine, Pinterest, Tumblr and Storify. Each of these platforms requires a new evaluation and approach."

collection of briefings, speeches and agency news. And the number of social media platforms continues to expand, with services such as Pinterest growing in popularity among government agencies this year.

Against that backdrop, agencies have started retaining and archiving social media. It's a challenging endeavor. They need to determine which communications must be preserved and then devise archival strategies for a still-evolving set of platforms.



Figure 18: FCW Article on the Next Generation of Archiving¹⁴

Maureen Pennock published a very comprehensive review of Web Archiving¹⁵ under the auspices of the Digital Preservation Coalition (DPC) in March of 2013. The review is published as part of the DPC's regular Technology Watch report series, which are intended as an advanced introduction to specific issues in digital preservation. This series is available on the DPC website¹⁶. Ms Pennock's review is an excellent introduction to the topic of web archiving, covering the motives for preserving web content and the technical challenges ranging from web crawler limitations to attempting to address long-term preservation issues such as format obsolescence and the complexities of interdependencies on the files that constitute a website. Ms Pennock's discusses the approaches of several projects which the DPC or its members are either involved in or are aware of, as well as many tools which can help with the specific problems of web archiving, such as DROID¹⁷ and PLATO¹⁸.

David Rosenthal is well known in the Digital Preservation community and had a blog post in July 2013¹⁹ about the use of URL shorteners on the internet which is another aspect of the social network that is worth citing as an example of the difficulties facing those seeking to preserve online content. URL shorteners are used by services such as Twitter to create short versions of very long URL's in order to keep the size of the tweet as small as possible. They are also commonly used in blogs and other forums. Taking the example of bit.ly, each time they shorten a URL, they write a re-direct into an Amazon S3 bucket so that a DNS redirect of the bit.ly URL can act as a backup plan in case their service ever goes down. The problem David refers to is that these URL lookup services are highly dependent on the service provider staying in business and that some sort of ESCROW, or out-of-band backup is needed in the event that the service provider is no longer around. Without consideration of these sorts of complexities, it's quite possible that any attempt to create an archive of social content on the internet will be populated with large numbers of orphaned and unresolvable URL's. Indeed, Maureen Pennock makes supporting comments when her

report speaks about much of content on the internet being transient by nature, for example news services and other outlets overwrite the oldest content with newer ones all the time resulting in users encountering http 404 file not found errors quite frequently.

As another example, in March 2013, Google announced the end of its Google Reader service²⁰. This service allows users to view blogs through an RSS feed. It was a timely reminder of how an online service that we know and perhaps come to depend on can disappear very quickly with no plan in place to capture the data contained within the service, its content can be lost forever.

Memento²¹ is a project funded by the US Library of Congress which has developed a browser plug-in (currently for Chrome only) which can allow a user to almost seamlessly browse the historical internet. The Memento plug-in makes it possible for a user to browse archived collections of the internet which are maintained and made available from groups such as the UK National Archives, various national Web Archives and the Internet Archive, to name a few. The Memento website provides an architectural overview of how this works which is shown below in Figure 19 and the site states *“the HTTP-based Memento framework bridges the present and past Web. It facilitates obtaining representations of prior states of a given resource by introducing datetime negotiation and TimeMaps. Datetime negotiation is a variation on content negotiation that leverages the given resource's URI and a user agent's preferred datetime. TimeMaps are lists that enumerate URIs of resources that encapsulate prior states of the given resource. The framework also facilitates recognizing a resource that encapsulates a frozen prior state of another resource.”*

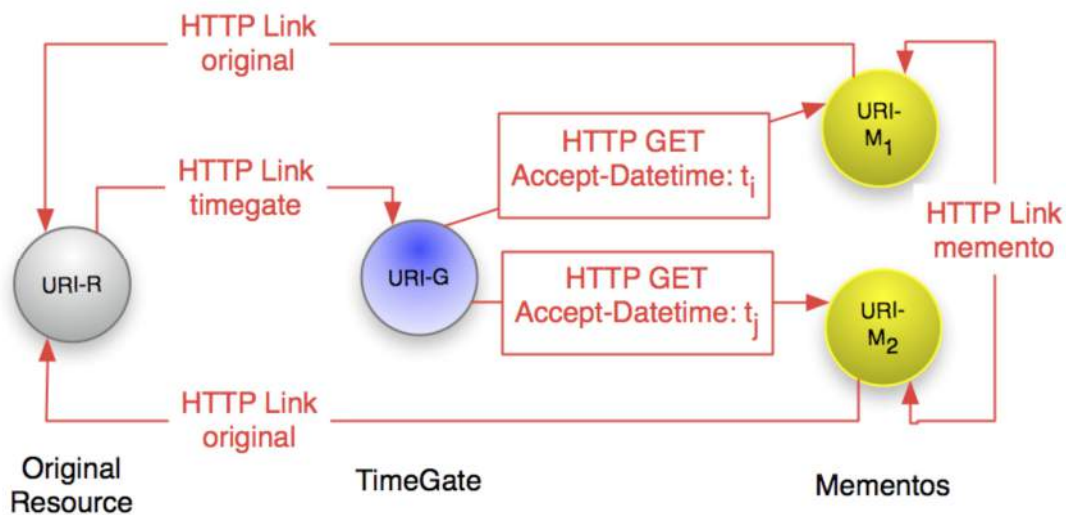


Figure 19: Architecture Overview of Memento Framework²²

Memento is interesting because it shows that there is a desire to access historically archived versions of the internet in an almost seamless way. The natural extension of this work is that it would grow to encapsulate the social network.

3.6 The Mobile Network

The mobile network is growing rapidly. As shown below in Figure 20, according to *Statistica*²³, mobile phones now account for 17% of total web usage and that figure excludes tablets which may be considered mobile devices by many. The growth in this traffic is staggering, in the 12 months from July 2012 to July 2013, this growth is just under 40% year-on-year which makes it one of the fastest growing ICT markets.

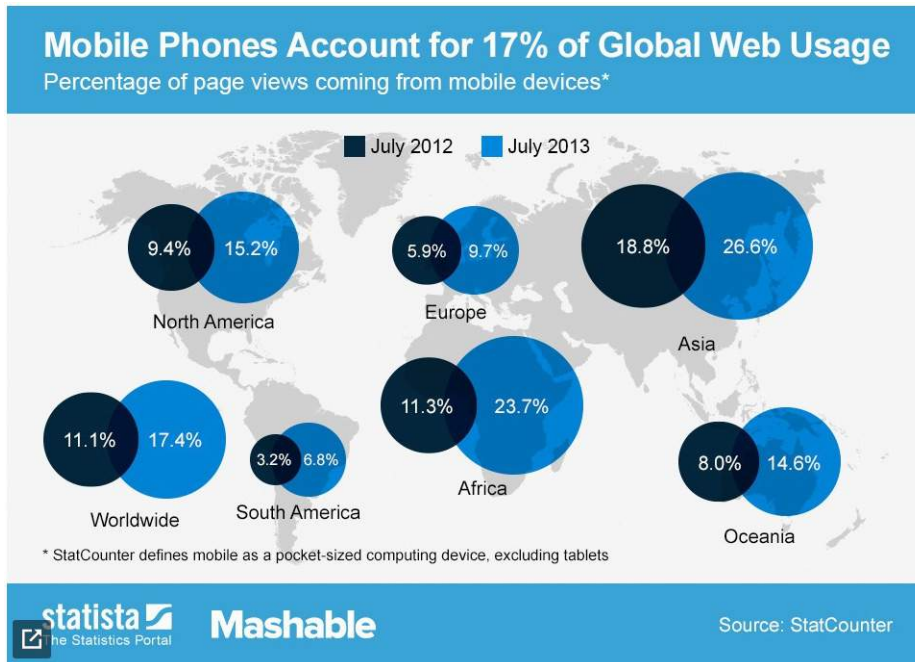


Figure 20: Mobile Global Web Usage²⁴

The use of mobile networks for internet access is significant for a number of reasons. First, it shows that computing is becoming more pervasive than ever before. We seem to want the internet at our fingertips all the time which is part of what is driving these figures. That in turn shows that the devices we are using to access the internet must change from the traditional PC which was large, heavy and typically fixed to one location in the house or office, to laptops and ultrabooks which were much better but still somewhat cumbersome to what we have today when we use smartphones and tablets. The second effect of this technology is that we are spending more of our lives online. That's significant because more of our footprint than ever before now exists in the essentially transient data of an increasingly digital existence. This section of the report will further explore this trend and its likely impact on Digital Preservation research. Figure 21 shows the mobile market trends. We see that users in China (to take just one geographic region) now access the internet through a smartphone more often than they do through a PC. This is happening right across the world and is correlated to the huge increases in global mobile network traffic generated by these devices. If we make the assumption that smartphones are a disruptive technology which will eventually completely replace traditional mobile phones, then the last graphic in Figure 21 shows that we still have the potential for huge growth in that market. And hence, internet pervasiveness will continue to grow in our daily lives for the foreseeable future, potentially changing the fabric of our society in, as yet unknown ways.

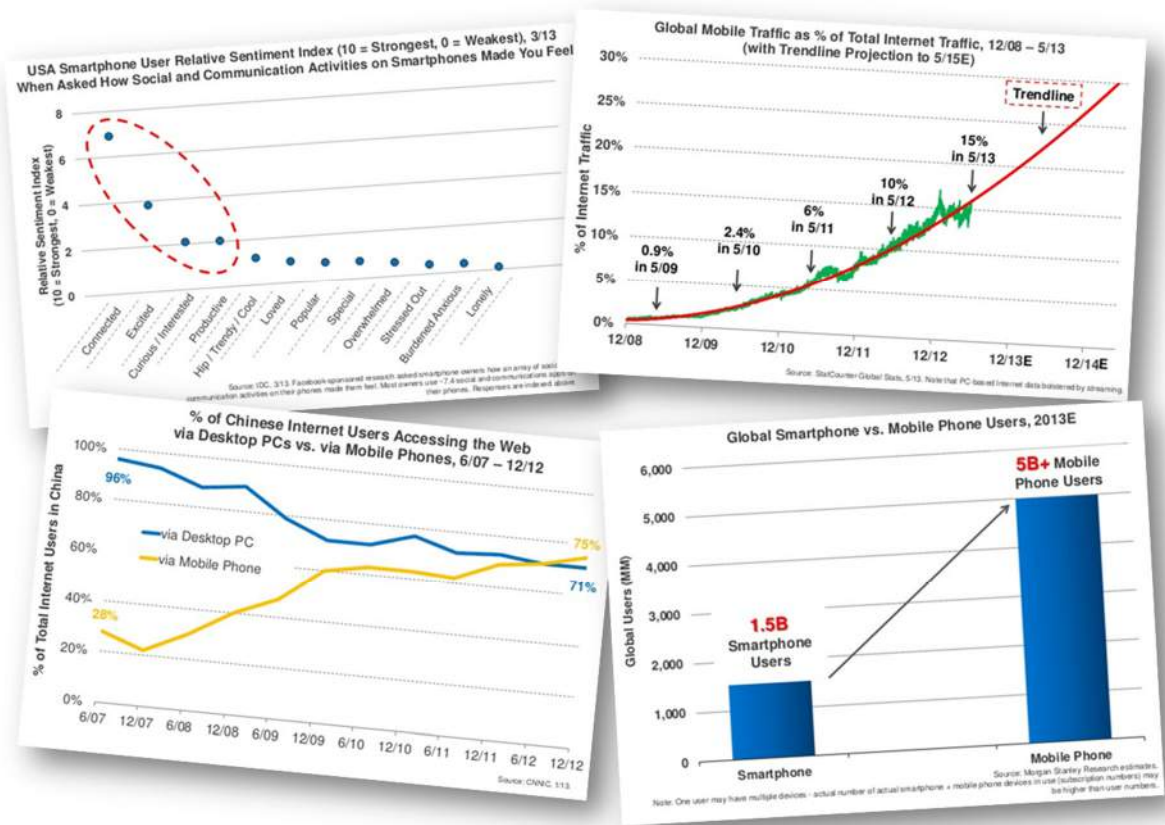


Figure 21: Mobile Market Research

(sources quoted in each, images produced by KPCB Internet Trends Report 2013)

Consider the effect this pervasiveness is having on the latest generations as they grow up today. World markets are much more closely aligned than any previous time in the last 100 years with Europe, the USA and China being the largest economies with a much less disparity between their GDP's as a percentage of global GDP as shown below in Figure 22. The world economy of the past was shaped by colonial expansion and then the waning power of those same colonial countries. The undiscovered country of the world is no longer to be found in the physical realm, but rather in the cyber one. Today's generations are re-imagining the world and their interactions and experiences online are driving that. Individuals such as Steve Jobs (Apple), Mark Zuckerberg (Facebook) and Larry Page/Sergey Brin (Google) are taking up the legacy of the microprocessor era and driving it forward in a way that could never have been foreseen before. They are changing the world through the realisation of their dreams made possible by today's technology. The implication for Digital Preservation research is that rapid technical evolution magnifies the problems of long term preservation. Not only does it create new devices and software environments but as a social movement, the momentum created behind it is accelerating as there are, of course, large potential markets for the companies and individuals who are successful with these undertakings.

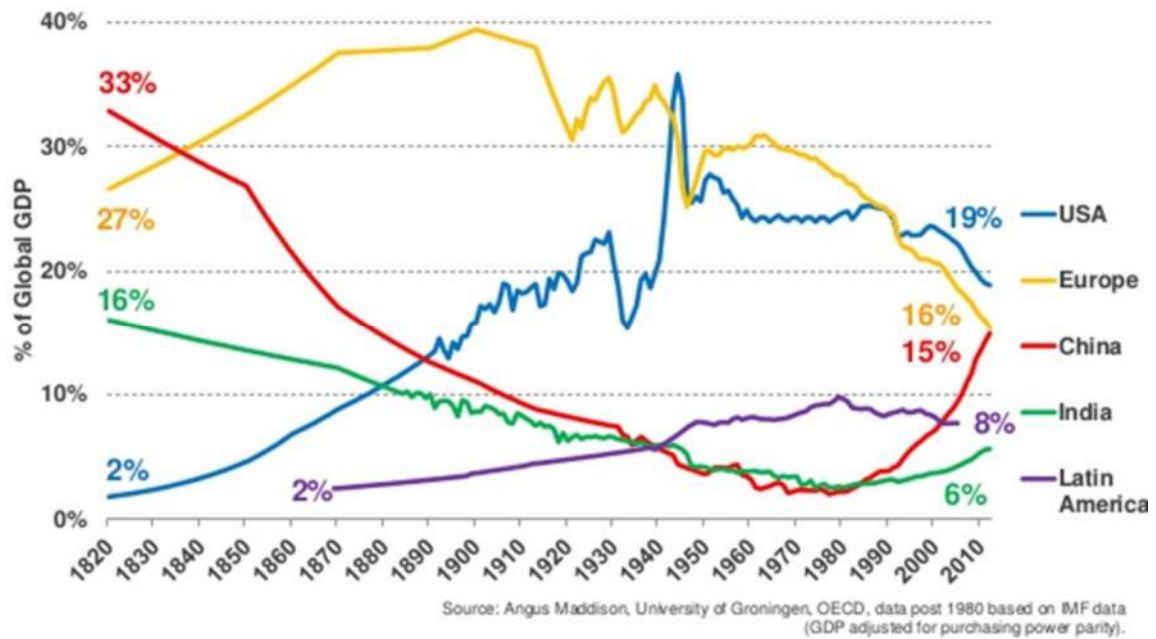


Figure 22: Percentage of Global GDP 1820 to 2012

(image sourced from KPCB Internet Trends 2013 report)

We need to think differently about these trends when we consider them from a Digital Preservation perspective. Figure 23 shows some interesting trends over time. Firstly, the rise of 'Wintel' in the 1980's to an almost dominant global position which lasted into the early 2000's. The term Wintel refers to the combination of Windows operating systems running on Intel architecture which became the cornerstone of the personal computing revolution. Before that period, there was a mix of gaming operating systems in the home as PC's had not become mainstream. The emergence of smartphones and tablets in the last 10+ years has seen the rise of the Android operating system with its easy to use touch-driven interface. The significance of this should be obvious to a digital preservationist; every technology appears to have a natural time cycle and it is crucial that efforts are made while the technology is still alive to preserve it. After-the-fact preservation becomes more challenging. This is because there is effectively a sliding time window where a critical mass of knowledge exists in the minds of software developers and hardware manufacturers about how to create, operate and support these systems. As Figure 23 also shows, a similar phenomenon can be seen across other technology spheres. In just eight years from 2005 to 2012, the rise of smartphones has turned the mobile operating system on its head. This has also affected browser usage. In just two years, we are seeing the effect of the change in our usage patterns. The browser change is interesting because many applications are web based and these tend to be very sensitive to particular versions of particular browsers, which in turn, will only run on certain operating systems, which again in turn, only run on specific devices or hardware. Change is usually good, but for Digital Preservation researchers, this rapid change presents huge challenges if we are to be successful in capturing digital representations of legacy systems.

Even programming languages are not immune from extinction. There has never been such a diversity of programming languages and platforms as there are today. Figure 23, for example shows the huge rise in JavaScript since mid-2009, which is almost certainly tied to the rise of NodeJS²⁵. The question is will we continue to see an increasing number of new languages developed or will they eventually converge to a smaller set which would be easier for entities such as ESCROW providers, or more specifically, their customers, to deal with. Preserving a programming language is not a trivial undertaking. It's not just a case of preserving the documentation of the semantics and constructs of the language. It extends to having the compilers to turn source code into executable instructions, it involves development tools and the environments in which those languages used to exist and it incorporates access to the countless numbers of libraries what are developed and maintained by a vibrant and growing open source community worldwide. Programming languages are alive and share many characteristics with spoken

languages in that they are constantly evolving and even taking on new functions, but the correct interpretations of their semantics can only be valid at a certain point in time.

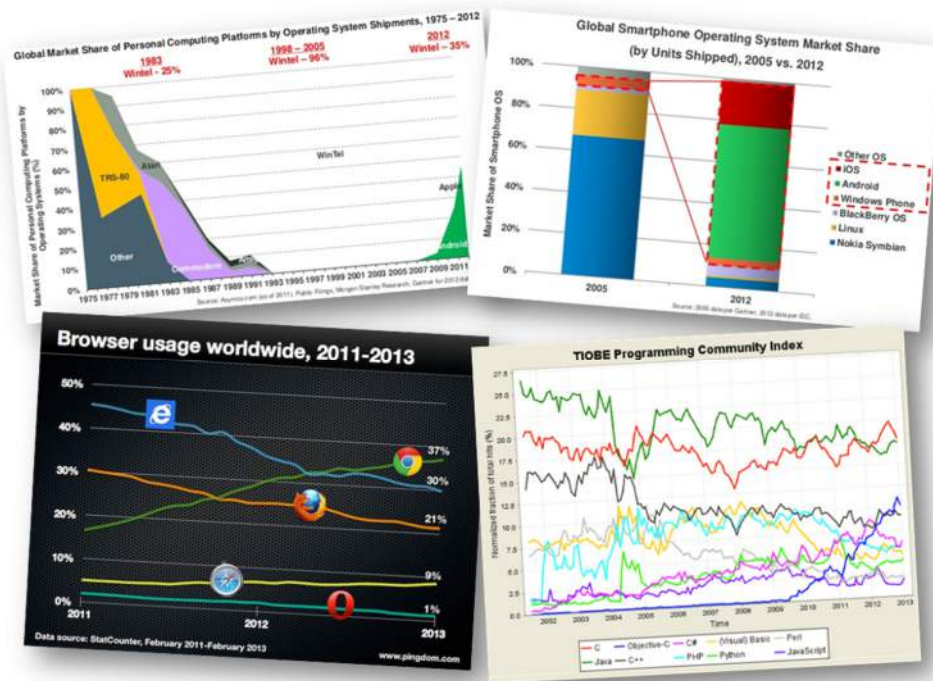


Figure 23: Wintel, Mobile OS, Browser and Language usage over time²⁶

3.7 Growth of Embedded Devices/Internet-of-Things

The intelligent systems market is growing. In 2012, this market is expected to reach more than 4 billion units and create \$2 trillion in revenue by 2015. This was according to International Data Corporation, a provider of market intelligence data and that by 2015, these systems will account for about one-third of all unit shipments of major electronic systems, compared with 19% in 2010. In the latest report²⁷ from IDC on this market, published in October 2013, the trends are accelerating. IDC now expects the number of connected autonomous devices to reach 30 billion by 2020 driven largely by intelligent systems which will be gathering data on behalf of their owners both end-consumers and enterprise. IDC now expects technology and services revenue from IoT to increase from USD\$4.8 trillion in 2012 to USD\$8.9 trillion in 2020 with a 7.9% CAGR (compound annual growth rate).

Figure 24, below shows the wide proliferation of IoT devices across industry at the current time. These devices are becoming far more widespread than is readily apparent to the casual onlooker.



Source: Beecham Research, *M2M/IoT sector map*, 2013.
 Graphic: Deloitte University Press | DUPress.com

Figure 24: IoT; Proliferation across industry²⁸

Figure 25 below illustrates the huge growth in M2M (machine to machine) or IoT devices predicted out to the end of the decade. While all available indicators show growth like this, one of the challenges that the market faces which is relevant to a digital preservation research perspective, is the lack of standards. As with many new technologies, priority is given to getting to the market first by many solution providers. This is understandable when all available market assessments are showing such potential growth in the coming years. However, this rush to market is a large driving force in adding to the complexity of digitally preserving business processes which contain IoT elements. Not only is there the possibility that some sort of functionally-equivalent, point-in-time emulator needs to be developed or preserved, but that the data formats and potentially also the network transports used by today's IoT devices are not necessarily based on mature standards and therefore subject to major revision in the future. We can only speculate about the lifecycles that the IoT devices themselves will have as these will vary on a use-case by use-case and device by device basis, but what is almost certain is that some will not survive in this competitive environment. They will either be rapidly replaced by newer, improved devices from the same supplier, in which case there is a better chance that the device will have some backward compatibility built-in, or they will simply disappear without a trace as the vendor either exits the market or goes out of business entirely. The relevance of this practice to digital preservation is important to grasp; essentially, the longer a technology is available, the more widespread its use is and the more standard based the technology is, the better it is from a digital preservation perspective. Once again, short lifecycles, rapid technological advancement and lack of standards are major challenges to successful, and complete, digital preservation as it is conceived in the TIMBUS project.

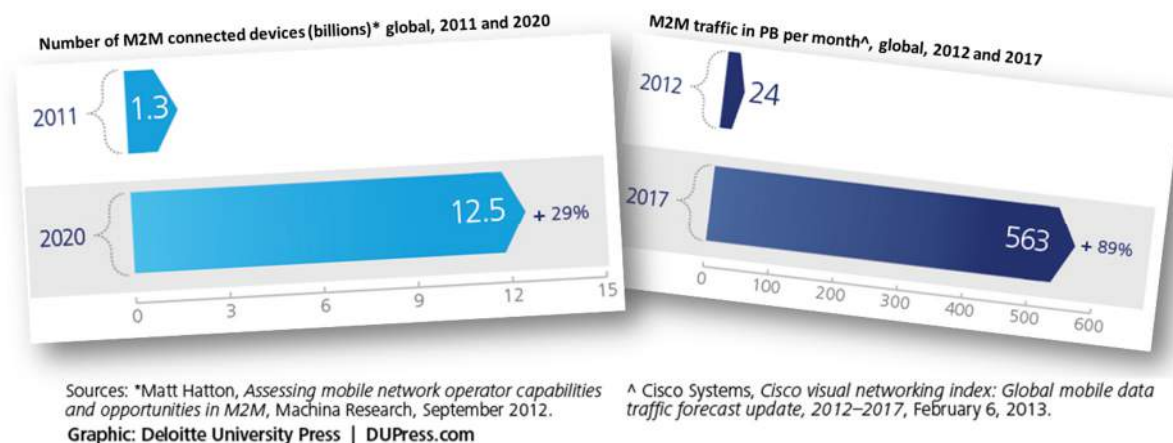


Figure 25: Number of M2M connected devices and M2M traffic²⁸

3.8 Conclusions

The digital universe is still growing rapidly. Richer content means growing volumes of network traffic. A recent internet article²⁹ by ARS Technica covered a story about Google and Microsoft joining in a push for a wider Wi-Fi spectrum. In that article, Cisco makes the following prediction: *"by 2017, Wi-Fi devices will power a majority of all Internet traffic. All of the devices you and your family use every day rely on Wi-Fi—and those demands are only increasing. And with the 'Internet of Things,' machines will need more and more of it to communicate with each other. But this growth may mean that getting on the Internet through your Wi-Fi connections will soon be like trying to drive in rush hour traffic on too narrow a road—frustrating and slow-moving."*

Underlying this growth, or perhaps driving it, are new usage patterns as people interact with technology in an increasing number of ways and through new devices. Never before has there been such diversity in devices and form factors available to technology consumers. The rate of change and adoption of these devices appears to be accelerating which exacerbates the challenges faced by digital preservation research. Wearables and the Internet of Things (IoT) are expected to grow rapidly. IoT in particular is discussed in an excellent Forbes article where Glen Martin interviews Jim Stogdill of O'Reilly's³⁰ in February 2014 titled *"How the Internet Of Things Is More Like The Industrial Revolution Than the Digital Revolution"*³¹. In it, Stogdill compares the aftermath of the Centennial Exposition of 1876 (Americas first World's Fair) to the point that IoT has reached today: *"It's really a watershed moment in technology and culture. We're at one of those tipping points of history again, where everything shifts to a different reality. That's what the 1876 exposition was all about"*. While the validity of such statements remains to be seen, there can be no doubt about the difficulty, and the increasing complexity of the task which a digital preservationist faces today. We have never been as dependent on technology as we are today and we have never had such a large penetration of technology into our daily lives as there is today. As we attempt to preserve technology solutions from today, the research community will struggle to keep pace with this rapidly and constantly changing landscape.

4 Global Archives

The continuing growth of data in the digital universe would naturally be expected to have an effect on global archives so it is no surprise to see that these are also experiencing rapid expansion. This section of the report augments this and has been augmented wherever more recent figures are available. All the sources for the various graphs and data points presented are quoted.

4.1 Worldwide Enterprise storage

Figure 26, below, shows IDC's forecast for storage capacity revenue and storage capacity shipped from 2008 out to 2017. There is strong growth forecast for each of these as demand for storage grows.

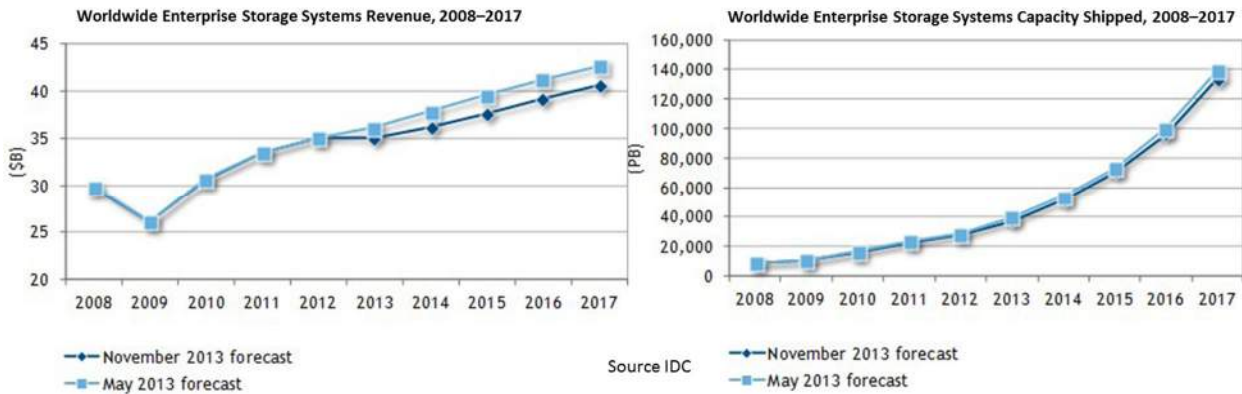


Figure 26: IDC Storage Capacity Revenue & Capacity Shipped

This demand is being fuelled by the growth that we are seeing in data in the digital universe as previously covered in section 3 of this deliverable.

However, some key market changes have happened since that ESG report took place and it is interesting to consider what is happening in the industry. In May 2012, IDC released a similar report entitled "Worldwide Enterprise Storage Systems 2012-2016 Forecast"³². As can be seen from Figure 27, below, the cost per GB for storage is reducing more than what were predicted in last year's figures. Although, as the BackBlaze graph shows, this is still not as low as before the Thailand flood crisis, the latest IDC figures, inset in Figure 27, show revenue out to 2017 is half of what was forecast just 12 months ago.

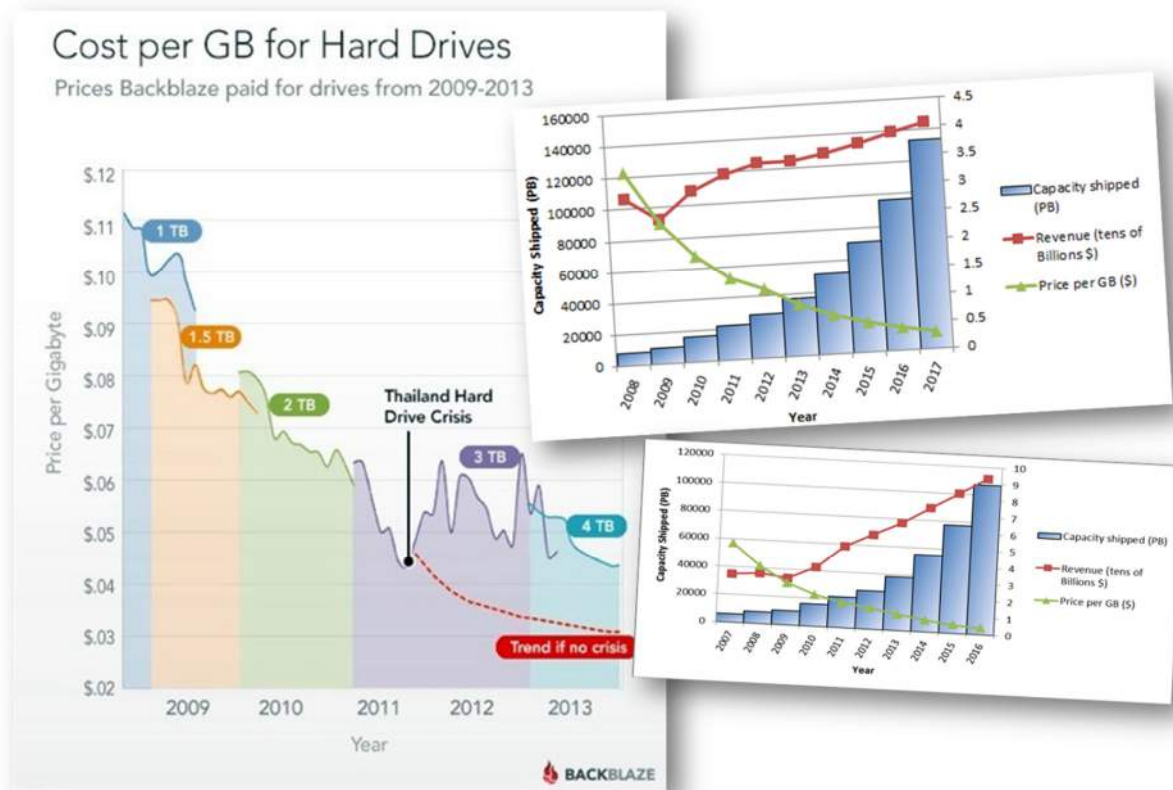


Figure 27: Cost per GB from BackBlaze and IDC³³

(inset graph is a composite of figures provided in table 9 on page 42 of that report for this 2012 and revised in 2013)

Demand has not reduced, so the explanation for the lower revenue is coming from projected reductions in storage costs of approximately 24% in CAGR (Compound Annual Growth Rates) terms. This is finally some good news for long-term archival solutions as the industry recovers from the Thailand flooding crisis, there would appear to be a reasonable prospect of a return to the days of cheaper storage solutions once more. However, it should still be noted that although the cost per gigabyte is falling again, this is not completely offsetting the gain in demand for storage. The figures still say that customers will be spending more overall to meet their storage requirements, but that spend is now increasing at a much slower rate than previously predicted and is in the order of under 20% out to 2017 versus the prediction last year of closer to 70-80% more cost to meet the demand.

4.2 The Software Landscape

The latest revenue predictions available from IDC relating to the storage software market are shown below in Figure 28. They forecast growth out to 2017. Interestingly, this is being led by Windows and Linux operating systems which are perhaps a reflection of these operating systems being heavily deployed across enterprise. Proportionally, revenue on Linux storage software is growing faster than Windows, but Microsoft operating systems are predicted, in these figures at least, to continue to hold a significant revenue lead. That may be a good metric for vendors of storage software on Windows platforms; another view point may equally argue that it is a worse situation for their customers who are perhaps paying more for services that are either cheaper or free on Linux or other open source platforms. From a digital preservation research perspective, the trend below is what is important rather than the scale or the exact figures forecast. The trend is predicting growth in software storage revenue meaning that it is reasonable to expect that there will be opportunities for those organisations in the market who can offer differentiated services over their competitors. There is no reason to think that those differentiations would not, at least in part, be based on enhanced support for long-term retention of data.

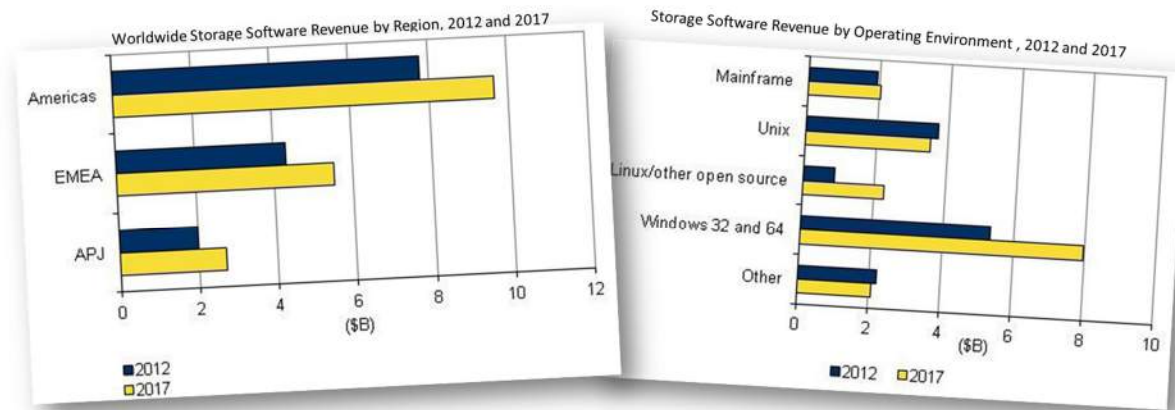


Figure 28: Storage Software Revenue³⁴

A 2014 report titled *IDC's Worldwide Cold Storage Ecosystem Taxonomy, 2014*³⁷, refers to some of the characteristics that IDC believes will be required for long-term archives. Namely, these are:

- **Open Standard Access:** This should include Open API access and metadata to support querying of the integrity of data in the archive. We know in the research community how important open standards are to long-term preservation solutions. A closed-source or proprietary solution is a large deterrent to many organisations that have seen the impact of vendor-lock in's previously. In the area of archival, those are even more difficult to resolve. This is perhaps why, IDC in another report titled *Worldwide Enterprise Storage Systems 2013–2017 Forecast: Customer Landscape Is Changing, Defining Demand for New Solutions*³⁵ talks about the re-emergence of 'do-it-yourself storage architectures'.
- **Media:** Cold storage systems should support per device power-off states and protection mechanisms against bit-rot. There is an opportunity here for solutions which optimise incoming and outgoing traffic to disks to maximise power savings by minimising the number of disks which need to be spun up to read or write data. Similar opportunities would exist for performance (i.e. IO) optimisations. This could be an area for future research in the preservation arena.
- **Data Security, Compliance and Durability:** Data needs to be secured, even at rest, so that only authorised users have access to it. Archival systems need to implement automated policy based data management capabilities. Many commercial offerings today are simply a deployment of some policy-based management on top of more traditional storage solutions. Lastly, it goes without saying that data loss needs to be guarded against. In this regard, there are many examples of implementations of multi-site solutions which are capable of eliminating the risk of a single outage, be that a hardware failure, a software failure or an environment impact of some other kind (flooding, fire, etc) causing data loss.

4.3 Tiered/Cold Storage

An IDC report titled *IDC's Worldwide Cold Storage Ecosystem Taxonomy, 2014*³⁷ was published in February 2014 and referred to the phrase 'Cold Storage'. This phrase has been around since at least 2010 when Dell published a white paper called "Object Storage: A Fresh Approach to Long-Term File Storage"³⁶. Figure 29, below, comes from the IDC report in 2014 and goes further to give a definition of cold storage. Essentially, this is an emerging term which is now being used in industry where previously the phrase *archive* might have been preferred. Individual deployment projects may use their own definitions, or perhaps introduced more granular definitions for their own clarity, but broadly speaking, all such terms which refer to the storage and management of inactive data on low-cost storage tiers, would fall under the umbrella of 'Cold Storage' according to the IDC definition.

“Cold storage solutions are the lowest tier of data storage solutions with a total cost that is lower than the residual or perceived business value of the (inactive or hardly active) data sets stored on them”

“Cold storage collectively refers to a set of services, applications systems, and media that support an operational state or delivery mode specifically designed for storing “cold,” or inactive, data — a state in which a deliberate trade-off is made in data organization schemes and data access times to provide significant capital and/or operational savings”

Figure 29: IDC Definition of ‘Cold Storage’³⁷

A tiered approach for data storage is nothing. A typical classification scheme for tiers in an enterprise is given in the same IDC report and shown below in Figure 30. The fundamental concept is based on observations around the Pareto principle, more commonly known as the 80/20 rule. This rule seems to also apply to archives with vendors such as EMC, among others, saying that 90 days after the creation of a file, only 20% of your files needs to be accessed frequently and the remaining 80% is either never accessed again or only accessed very infrequently. IT service providers can take advantage of this and optimise the cost of storage solutions by implementing tiered storage infrastructures which automatically move data down to lower cost tiers over time if that data is no longer being regularly accessed. The trade-off in return for lower cost, is that the data consumer must also accept a lower performance level. This is achieved in practice through the use of less expensive, typically slower and typically higher capacity disks or tapes at the colder tiers. The hot tiers would involve the deployment of the highest speed disks, Storage Area Networks (SANs) and Network Attached Storage (NAS) which are generally more costly to provision on a dollar/euro per gigabyte scale.

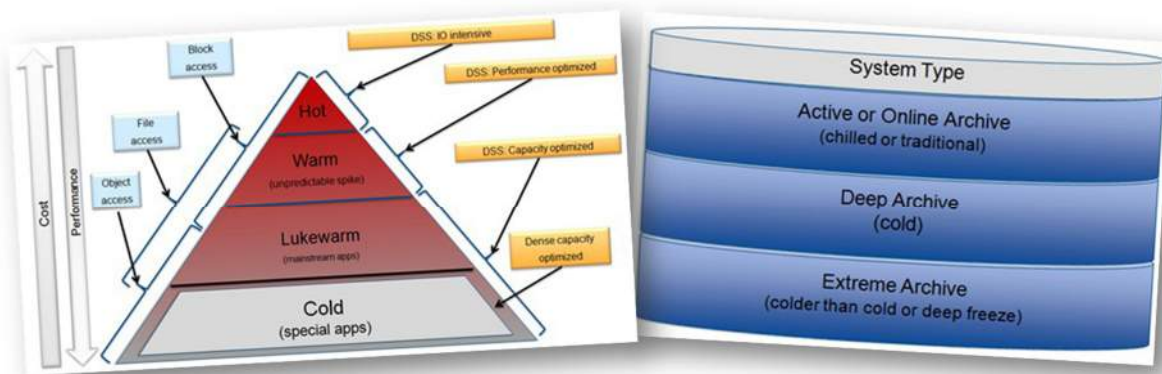


Figure 30: An Enterprise Classification Scheme for Cold Storage³⁷

Figure 31, below, also from the same IDC report gives a better understanding of the differences between the media used at the various layers of a tiered-storage solution. As can be seen, the seek time, or total retrieval time increases at the colder tiers.

| System | Media Used | Characteristics |
|--------------------------|--|--|
| Online or active archive | Flash | <ul style="list-style-type: none"> ▪ Mainly random read/write access ▪ Mainly eMLC Flash |
| | Capacity-optimized HDD | <ul style="list-style-type: none"> ▪ Mainly random read/write access ▪ Mainly 7,200rpm but can be slower depending on the data transfer rate expectations for a given application ▪ Data retrieval in <1 second when HDD is idle |
| Deep archive | Flash | <ul style="list-style-type: none"> ▪ Mainly random read/write access ▪ Mainly eMLC Flash |
| | Capacity-optimized HDD | <ul style="list-style-type: none"> ▪ Mainly sequential read/write access ▪ Less than 7,200rpm — sufficient for data transfer rate expectations ▪ Data retrieval in ~30–60 seconds |
| Extreme archive | Capacity-optimized HDD | <ul style="list-style-type: none"> ▪ Mainly sequential read/write access ▪ Less than 7,200rpm sufficient for data transfer rate expectations ▪ Data retrieval in ~30–60 seconds |
| | Traditional magnetic tape or optical media based | <ul style="list-style-type: none"> ▪ WORM availability in hardware (for compliance) ▪ Sequential read/write access ▪ Data retrieval in >60 seconds |

Figure 31: Characteristics of Media used in Archives³⁷
(source: IDC)

The take-away point for Digital Preservation researchers is not the fundamentals of tiered-storage, which is not a new concept, but rather the emergence of this new nomenclature in industry. Large divergences can be observed in the language used by industry, government and memory institutions that may essentially be talking about the same concepts and trying to address similar challenges, but are on different wavelengths when it comes to speaking to one another to share common learnings. This is of course merely one example which is a symptom of cultural differences between these organisations due to their foundation, growth, goals and ambitions.

4.4 Archival problems facing organisations

Julia Lockner of Informatica³⁸ released a series of blogs on this topic in April 2013³⁹. These were published after Informatica hosted a webinar on Enterprise Data Archiving and surveyed the 600 attendees. This section of the deliverable uses their findings to draw some conclusions about the archival problems facing organisations today in order to arm Digital Preservation researchers and individuals tasked with archival roles in industry with updated and pertinent information we could find on this topic.

Figure 32, below, shows the responses from attendees about the largest pain points they are feeling due to the explosion in data. Over 60% say that costs and the increasing effort spent on maintenance are the biggest factors, while 18%, almost 1 in 5, say that data discovery is challenging. Interestingly, compliance concerns are low. This may be because organisations feel they have this aspect of their business under control or else they simply have much larger problems to solve first. The survey didn't offer any data to support either conclusion. The Informatica data lists data warehousing and reporting environments as by

far the fastest growing category of structured data which organisations are dealing with. When it comes to unstructured data, unsurprisingly, e-mail and office documents appear as the largest growing data types but interestingly we see images, video, social networking and mobile/call record data in the top six which is a reflection of the increasing visibility these data types are receiving. Informatica’s report listed databases as a type of unstructured data in their survey, which may be arguable as the more widely held opinion would be that these would constitute structured data. Leaving that consideration aside, the data would point to the requirement for organisations to plan for data growth in a deliberate manner and to develop strategies to cope with their individual circumstances. This is really a necessity from a budget perspective.

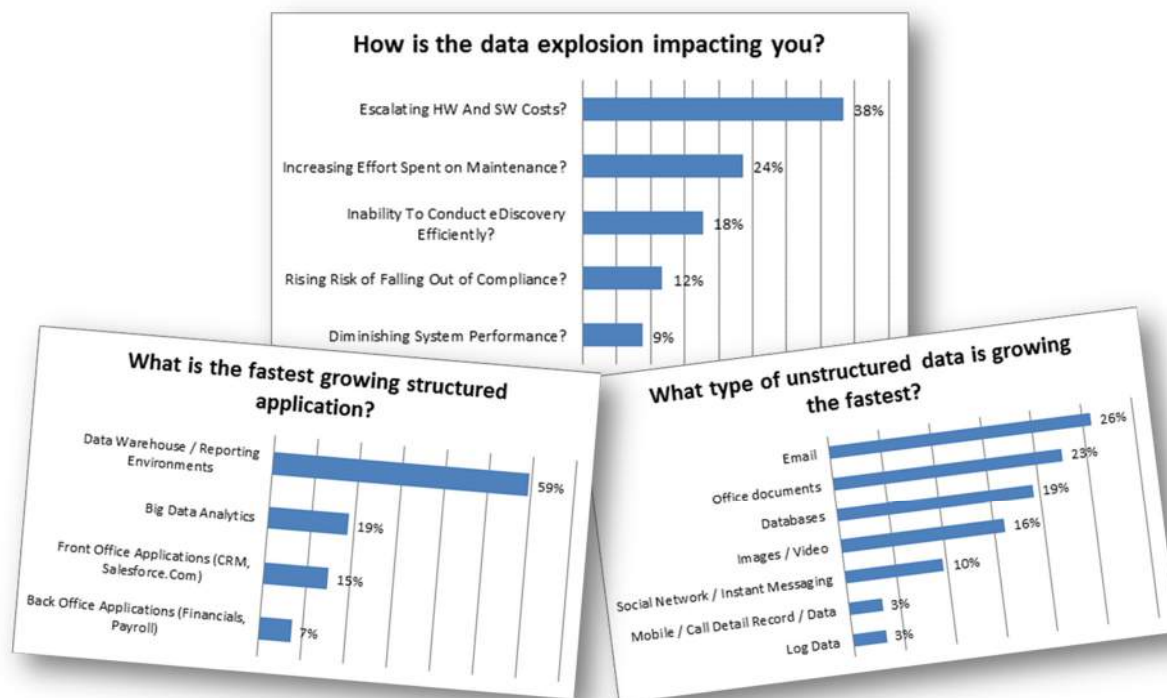


Figure 32: Responses on dealing with Data Growth
(source Informatica)

Informatica also surveyed their attendees about data archival. Figure 33, below shows these. Just under two thirds of respondents say their organisations are actively archiving data. Unfortunately, the definition of what actually constitutes archiving is open to whatever interpretation the individual respondents put on it, so it is not clear from these figures how many of those are for long term retention versus medium-term, versus perhaps, even just for disaster recovery purposes. From the TIMBUS perspective, the consortium would feel there is always a need to define exactly what is meant by *archival* as there are different interpretations of that it means to different individuals and organisations. The only insight offered by the figures is in the area of benefits which organisations expect to see from archival. Improving application performance and regulatory compliance are seen as the largest reasons to perform archival. Arguably, data centre consolidation, streamlining of IT operations and reduction of infrastructure costs are really different strands of cost management and if grouped together, this category would actually be the largest. If this interpretation is accepted, then cost management and application performance are seen to be a more important benefit than regulatory compliance.

Section 4.5 of this report will discuss using the cloud for archival but the topic is introduced here as part of Informatica’s blog. Again, the question is not as precise as we would like it to be as it does not make it clear if respondents are already using cloud services for archival or simply considering it. The conclusion this report draws from that question is that there is a small, but significant percentage of organisations strongly considering using cloud for archival, up to perhaps 25% and this number may be as high as 51% if the figure who responded with ‘maybe’ are taken into consideration. If we again make the reasonable

assumption that ‘consideration’ means an organisation feels archiving in the cloud is an option for them but they have concerns about doing it, then the survey shows a healthy level debate on the issue of using cloud for archival. Based on market assessments to date, it appears that such a move is highly risky, particularly for sensitive data or data which needs to be archived for a long period of time. Cloud providers are poorly equipped to meet the requirements of long-term-preservation and they are not cheap despite lower cost options such as Amazons Glacier service which undercuts their S3 storage service in pricing terms and is aimed at the archival market.

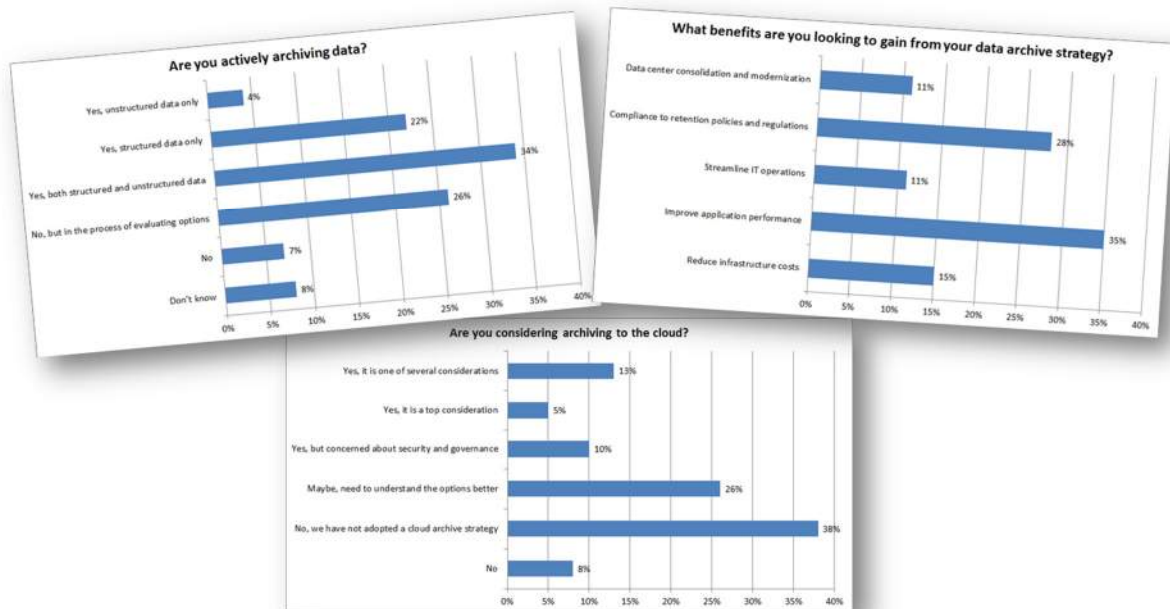


Figure 33: Responses on Archiving
(source: Informatica)

4.5 Using Cloud for Archiving & Digital Preservation

Figure 34 below highlights the top 6 barriers to adopting the cloud, as identified in a 2010 IBM report on Global IT Risks. These concerns still exist today and it is reasonable to assume these reasons underlie some of the considerations that the respondents to the Informatica survey had on using cloud for archival in section 4.4 of this report. Despite these concerns, it appears that there are a growing number of cloud services that started to come online in 2012 to support archival for those organisations who wish to entrust their archival requirements to an external cloud provider.

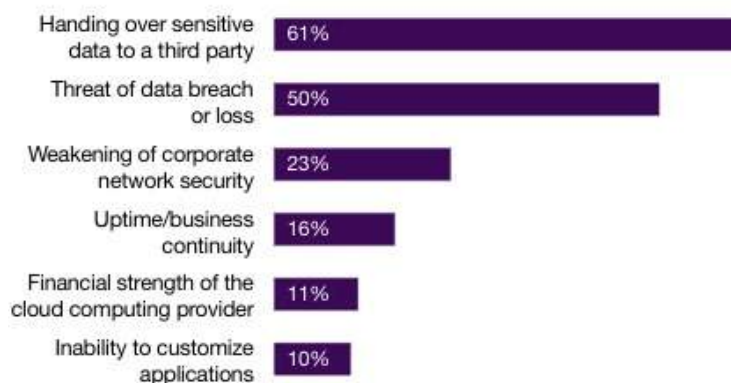


Figure 34: IBM Risk Global IT Risk survey 2010⁴⁰

Michael Peterson a consultant, Chief Strategist to SNIA has published on the LTDP (Long term digital preservation reference model) and he suggested some thought providing requirements for preserving information in the cloud which this deliverable referred to in its last iteration: ⁴¹

- Information must be Organised, Identified, and Indexed
- Complete, Authentic and Valid
- Catalogued (including metadata & required manifest) and Tracked
- Transformed to a standard format
- Transferred reliably and verified (non-repudiation)
- Tamper proof
- Confidential, secure, protected from change
- Digitally Audited and protected from damage or loss
- Accesses Logged
- Direct Access
- Self-validating
- Migratable (physically and logically)
- Portable (independent of the physical storage) and interoperable
- Useable and accessible
- Retention management including Deletion capabilities

As we go through 2014, this debate is continuing. In one sense, it is a subset of the larger, but related discussions around cloud (in the sense of external hosting) versus internal hosting of services. The market that the cloud service providers cater to has been one which is the least concerned with privacy and assurances about Service Level Agreements (SLA). This has allowed cloud providers to build up a reasonable sized, and very profitable market while at the same time, putting very minimal, if any, effort into addressing concerns from that segment of the market which is unwilling to use external cloud providers due to the lack of SLAs and quality of service guarantees. We have seen previously in section 3.5 about how the Google Reader service was shutdown. This can also happen to cloud providers. For example the Register reported an article⁴² on Nirvanix, who were a cloud storage provider, going out of business. According to that article, on September 18th 2013, Nirvanix customers were told that they had until the end of September to get all of their data out of the cloud or risk losing it despite a plan being spun up to save the service and migrate its customers and data to a sister company. David Rosenthal picks up on this point in one of his blogs⁴³ posted October 1st 2013. Working with some approximate figures for estimated data volumes stored in Amazon S3, the likely available bandwidth for copying data out of Amazon and into various customer sites and the problem of sourcing storage disks at short notice, he compares Amazon S3 to now being like banks in that they are “too big to fail”. The potential knock-on effect to industry and government of being given a 13 day ultimatum to move data from the cloud, or face losing it from a service as large as Amazon, would cause huge disruption. David Rosenthal estimates that to be in the region of over USD\$1 Billion and potentially, the government would have to intervene. Bearing that example in mind, it is hardly surprising to find that organisations are giving very careful consideration before deciding to host data externally; particularly in the case of long-term archives.

Truman Technologies published a blog article⁴⁴ in February 2013 which discussed the top 5 questions to ask before outsourcing archival to a cloud vendor. The five questions raised in the blog, and the thoughts of the TIMBUS consortium on each of the points raised in the blog are given below in Table 2.

Table 2: Top 5 questions to ask before outsourcing archival to a cloud vendor

| Question | TIMBUS assessment |
|--|---|
| <i>What happens to my digital assets if the vendor goes out of business or ends the service?</i> | We have seen what happened to Nirvanix when they went out of business. We know that several storage providers are brokers who offer services on top of a tier 1 provider such as Amazon or Google. The relationships which these organisations have with the underlying provider are not necessarily visible to the end consumer. SLA's, if |

| | |
|---|--|
| | <p>present at all, are difficult to measure and if terms are breached, the financial compensation at best will amount to a partial refund of the subscription cost of the service. There is no allowance made for the value of data beyond that.</p> <p>Cloud providers have been hesitant to provide open API's and interoperability between cloud providers is still not very advanced. Vendor lock-in is a distinct possibility and the data owner has very little influence on the providers' charges.</p> |
| <p><i>Does the cloud vendor understand the requirements of digital preservation and archiving versus digital storage?</i></p> | <p>The Minnesota History Society's assessment of cloud vendors carried out in April of 2013 was referred to in this deliverable. This is the latest independent assessment available to us and its conclusions are that very few cloud providers understand the requirements of long-term preservation. Cloud providers are essentially offering a "data-dump" service to customers with very few guarantees. There are some niche players who do understand the problem domain better than others, but in general terms, the largest players in the cloud storage market are not catering for long-term archival needs and that takes into account services such as Amazon Glacier which is aimed at archival. The reason is they don't have to. The part of the market that cares about long-term preservation is still small enough that the larger vendors feel it can be safely bypassed for the moment while they concentrate on winning as much of the low-hanging global storage business as possible.</p> |
| <p><i>How much preservation management and workflow is handled or aided?</i></p> | <p>The OAIS model defines the basic steps in correctly preserving a digital artefact and David Rosenthal in a recent blog⁴⁵, stated that "<i>the research into the historical costs of digital preservation can be summarized by the following rule of thumb. Ingest takes about half, preservation (mainly storage) takes about one-third, and access about one-sixth of the total</i>". The ingest and storage are expensive processes which must generate (or at least support) metadata storage, fixity checking, replicas, potentially encryption/access controls, and so on. If a cloud provider is simply just providing a place to dump data and read it back, then they are not necessarily saving the organisation as much as they could be.</p> |
| <p><i>In what location will my archive collection be stored, and who owns my data and copyright?</i></p> | <p>These issues would apply to any storage of data in the cloud. It should be possible to know which geographies your organisations data resides in, if not the data centre itself. This may be of concern if the data is going to be held in a different legal jurisdiction. Cloud providers, for cost or performance optimisation purposes may have the right to move this location without your prior consent or even your knowledge. It may be advisable to find out if any legal assurances are given about this.</p> <p>Copyright access to data can often differ from the assumptions that users make about it. By storing data on a providers cloud service, your organisation may be giving up certain rights on how that cloud provider can use your data for their own purposes. It is best to take legal advice and check terms and conditions through a lawyer. What position is your organisation in if there are no details about access rights to data in the terms and conditions? This could potentially leave a service provider free to mine your data for whatever purposes they see fit.</p> |

| | |
|---|---|
| <p><i>How do I (easily) get my thousands of terabytes of digital content into – and out of – the cloud?</i></p> | <p>Sometimes a vendor may support an organisation sending in tapes or disks full of data to speed up the process of getting the data onto the vendor’s service. Network bandwidth can be a limiting factor for large data volumes in terms of time and costly to provide bandwidth. Cloud vendors also have metered tariffs for data access; if for example, your organisation wanted to make a digital collection accessible to the public and it proved very popular, there could be raising costs in continuing to host the service as more users access it.</p> <p>If a vendor goes out of business, will it even be possible to get your data back in a reasonable period of time?</p> |
|---|---|

4.6 Conclusions

Growth in global archives is continuing and increasingly cloud providers are attempting to lure this potentially lucrative business away from in-house data silos and on to their cloud storage platforms. This is clearly a much easier proposition for certain types of data and certain use cases than it is for archives. However, we should not discount the role of cloud providers entirely. A hybrid cloud approach may well make sense in certain situations. In this scenario, a cloud provider may keep one or more copies of the data for your organisation on their infrastructure while you also retain a copy. This may be beneficial as a protection against natural disasters or other service outages.

5 Influence of Emerging Technologies

In general, Big-data refers to data sets that are so large that they pose problems for traditional relational databases, or traditional tools and applications to process. A common example of a big-data problem would be an internet search engine because it combines three aspects that must be present for a task to qualify as a big-data problem. These are known as the three V's, namely these are Volume, Velocity and Variety and are explained well in a wired.com⁴⁶ article shown below in Figure 35 by Chris Taylor. Volume refers to the amount of data which needs to be processed. Today, big-data sets are many terabytes in size and petabyte scale data-sets are becoming more frequent. Velocity refers to the speed that the query needs to be executed in. Variety refers to the types of data to be queried which are no longer just rows and columns of text in a relational database but now include videos and images which are not so easy to perform contextual queries against.

WIRED GEAR SCIENCE ENTERTAINMENT BUSINESS SECURITY DESIGN OPINION

INNOVATION INSIGHTS community content featured blog

Share 37
Tweet 156
+1 66
Pin 87

Three Enormous Problems Big Data Tech Solves

BY CHRIS TAYLOR, TIBCO 08.01.13 2:32 PM

BIG DATA?

VOLUME
Large amounts of data.
Needs to be analyzed quickly.

VELOCITY
Needs to be analyzed quickly.

VARIETY
Different types of structured and unstructured data.

Worldwide IP traffic will quadruple by 2015.

By 2015, nearly 3 billion people will be online, pushing the data created and shared to nearly 8 zettabytes.

There's no time to intervene. This is big data's velocity. All of that digital data creates massive historical records but also rich streams of information that are flowing constantly. When we take the patterns discovered in historical information and compare it to everything happening right now, we can either make better things happen or prevent the worst. This is revenue generating and life saving and all of the other wonderful things we hear about, but only if we have the systems in place to see it happening in the moment and do something about it. We can't afford enough human watchers to do this, so the development of big data systems is the only way to get to better things when the data gives humans insufficient time to intervene.

Variation creates instability. This is big data's variety. Data was once defined by what we could store and relate in tables of columns and rows. A world that's digitized ignores those boundaries and is instead full of both structured and unstructured data. That creates a very big problem for systems that were built upon the old definition, which comprise just about everything around us. Suddenly, there's data available that can't be read or generated by a database. We either ignore that information or it ends up in places and formats that are unreadable to older systems. Gone is the ability to correlate unstructured information with that vast historical (but highly structured) data. When we can't analyze and correlate well, we introduce instability into our world. We're missing the big picture unless we build systems that are flexible and don't require reprogramming the logic for every unexpected (and there will be many) change.

There you have it... The underlying reasons that big data matters and isn't just hype (though there's plenty of that). The digitization, lack of time for intervention and instability that big data creates leads us to develop whole new ways of managing information that go well beyond Hadoop and distributed computing. It's why big data presents such enormous challenge and opportunity for software vendors and their customers, but only if these three challenges are the drivers and not opportunism.

I'd love to get your feedback.

Thanks to TIBCO CTO Matt Quinn for the ideas in this piece.

Chris Taylor is a marketing executive with TIBCO, and cofounder of Successful Workplace.

Figure 35: Wired.com article explaining the 3 V's of Big-Data

The internet search engine problem is a good example because the user expects an answer to their query in a fraction of a second, but the search engine provider needs to trawl hundreds of millions of websites globally, in different languages and across different structured and unstructured data sources. Clearly this is not a task that is executed at run-time or we would be sitting around for hours waiting for a result from a search engine. Therefore, internet search engines employ a series of steps to tackle the problem as shown below in Figure 36.

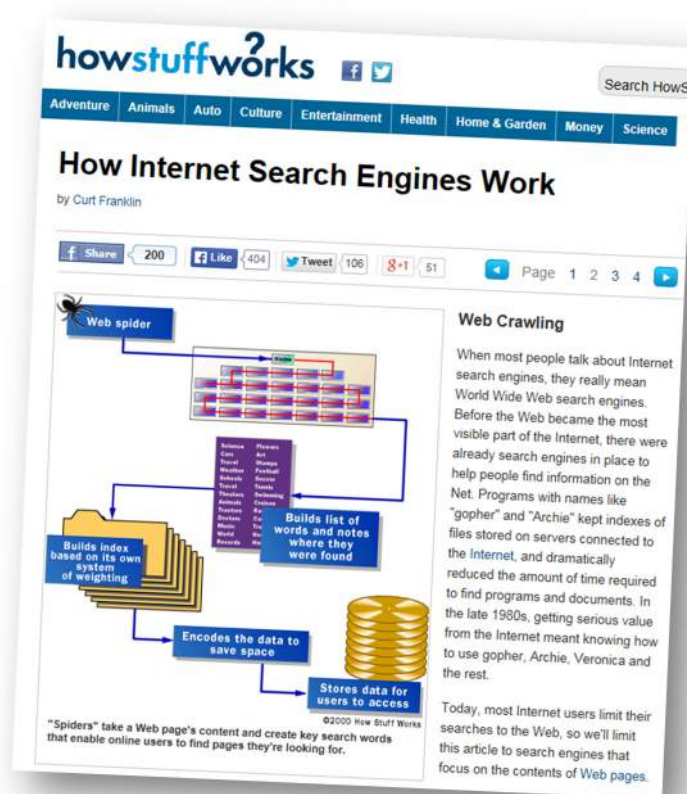


Figure 36: How Internet Search Engines Work
(source: howstuffworks.com)

Therefore, as we talk about big-data, we should bear in mind the three V's as the criteria that must be satisfied in order to have a big data problem.

5.1 Intel Distribution of Hadoop, SAP HANA

Intel and SAP both host websites dedicated to the topic of big-data. Both of these sites have predictably strong marketing aspects to them and are pitched as introductions to big-data. Intel's is called *Big Data Analytics Begins with Intel*⁴⁷ and SAP's is titled *HANA. Not just a database but a whole new approach to data*⁴⁸ based on how their in-memory database, HANA can provide a whole new approach to big data. Both Intel and SAP are actively targeting different aspects of the Big-Data market and have internal business divisions dedicated solely to producing products, services and solutions that meet the needs of these customers.

Intel's play in the market is centred around Intel's Distribution of Hadoop (IDH) offering while SAP's as previously mentioned is focused on HANA. Intel and SAP have actually teamed up on optimising these two platforms to work together as shown in some of their marketing slides below in Figure 37.



Figure 37: The Intel-HANA Story⁴⁹

These big data offerings are enabling tools upon which a specific application or service must be developed and built upon. IDH contains a management console called Intel Manager whose architecture is shown below in Figure 38. Likewise, HANA is a platform upon which solutions can be built, rather than a platform which provides them by default. The Intel Manager installs and configures components such as Hadoop, HBase, HDFS, Oozie, Mahout, Hive and zookeeper.

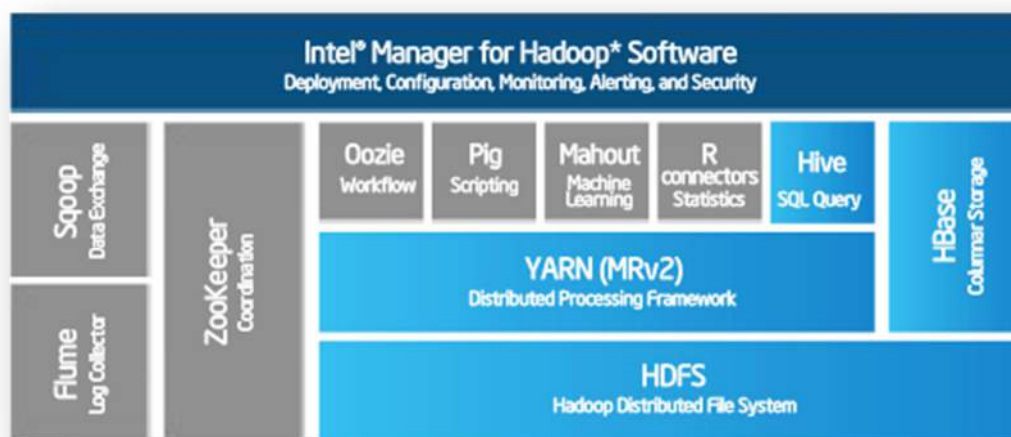


Figure 38: IDH Architecture for Intel Manager

(source: Intel)

5.1.1 Real-World Messaging Implementations

On December 9th 2013, Netflix publicly detailed a system they call Suro⁵⁰ which acts as their back end data pipeline. Suro uses Apache Kafka, Hadoop, HBase and Storm. Netflix published some pretty impressive numbers claiming that their architecture can process around 80 billion events per day. The Suro architecture is shown below in Figure 39. Also shown in Figure 39 is the Cloudera Impala⁵¹ architecture. On December 15th 2013, Amazon announced it was adding support for SQL-like queries on its AWS Elastic

Map Reduce service based on this technology from Cloudera. This architecture is also based on a Hadoop backend. But again, they show that all these big-data technologies are simply tools which can enable more sophisticated solutions to be built on top of them. But, as toolsets, they do help with problems involving massive data volumes and high computational loads.

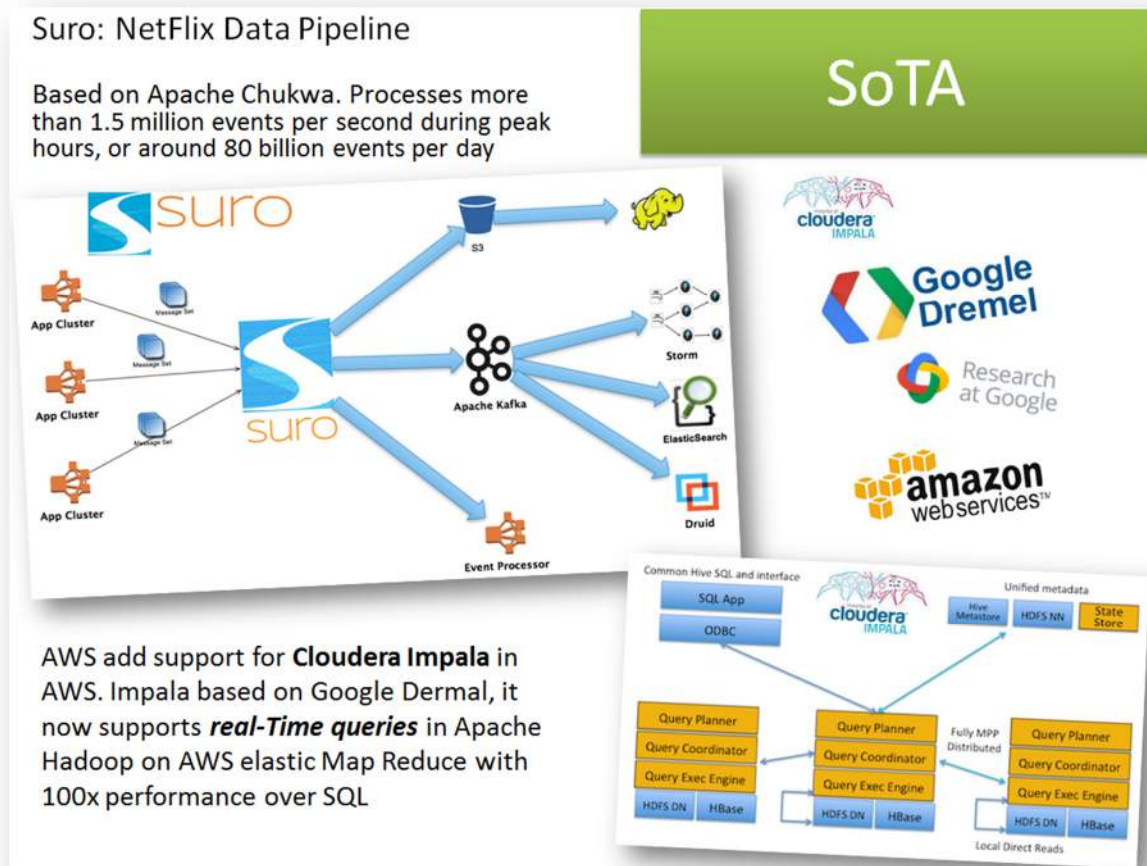


Figure 39: Other State-of-The-Art (SoTA) Implementations⁵⁰

5.2 Emerging Big Data Technologies

There is of course plenty of research taking place into emerging big data technologies which span spheres of computer science ranging from machine learning and reasoning to social network analysis, and semantic search. This section of the deliverable will consider some of these areas of emerging technologies and their applicability. Our research in this area was aided by an EPAC presentation on the potential of using such tools for anti-corruption⁵² purposes given by Mitja Jermol who is head of the Centre for Knowledge Transfer in Jozef Stefan Institute in Slovenia. The information about the capabilities of the tools in this section is taken from their official websites and hence sources for any information in this section are referenced and the material detailed here is firstly to inform the reader what these technologies and tools are and secondly to consider their applicability to TIMBUS-like scenarios.

5.2.1 Cyc & OpenCyc

Cyc⁵³ is an Artificial Intelligence (AI) project which has been running since 1984 with an objective of creating an ontology which is so detailed that it is capable of enabling an AI application to perform human reasoning. It is an ambitious undertaking to essentially create a huge database of common sense rules and knowledge, however Cycorp Incorporated who run the project have been very successful. Cycorp have their European headquarters in Ljubljana as detailed on their homepage shown below in Figure 40. They specialise in using their ontology and knowledgebase as a starting point upon which to build other services and applications rather than beginning with these tasks from scratch. They are also involved in

an FP7 project called LarkC⁵⁴ which has the ambitious goal of developing “a platform for massive distributed incomplete reasoning that will remove the scalability barriers of currently existing reasoning systems for the Semantic Web.” There is an open source version of this ontology called OpenCyc which researchers and developers are encouraged to try as an alternative to designing their own ontology. They also work with applying their expertise to the challenges of natural language processing which have proven to be a difficult obstacle for computer scientists.

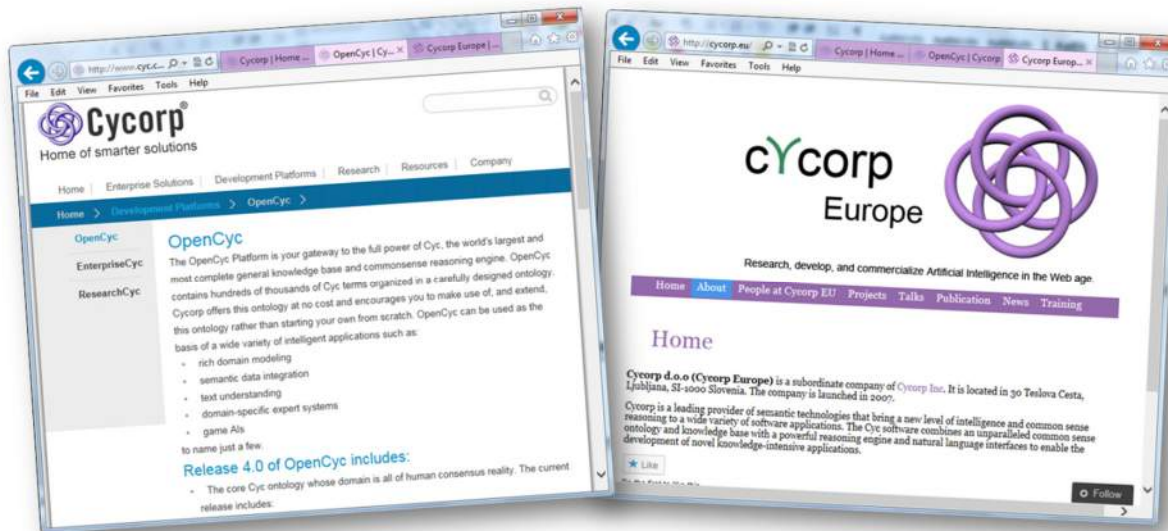


Figure 40: Cycorp and Cycorp Europe Homepages

5.2.2 COLIBRI, Apache Jena, OntoGen and OntoBridge

This section discusses a number of inter-related semantic ontology tools as shown below in Figure 41.

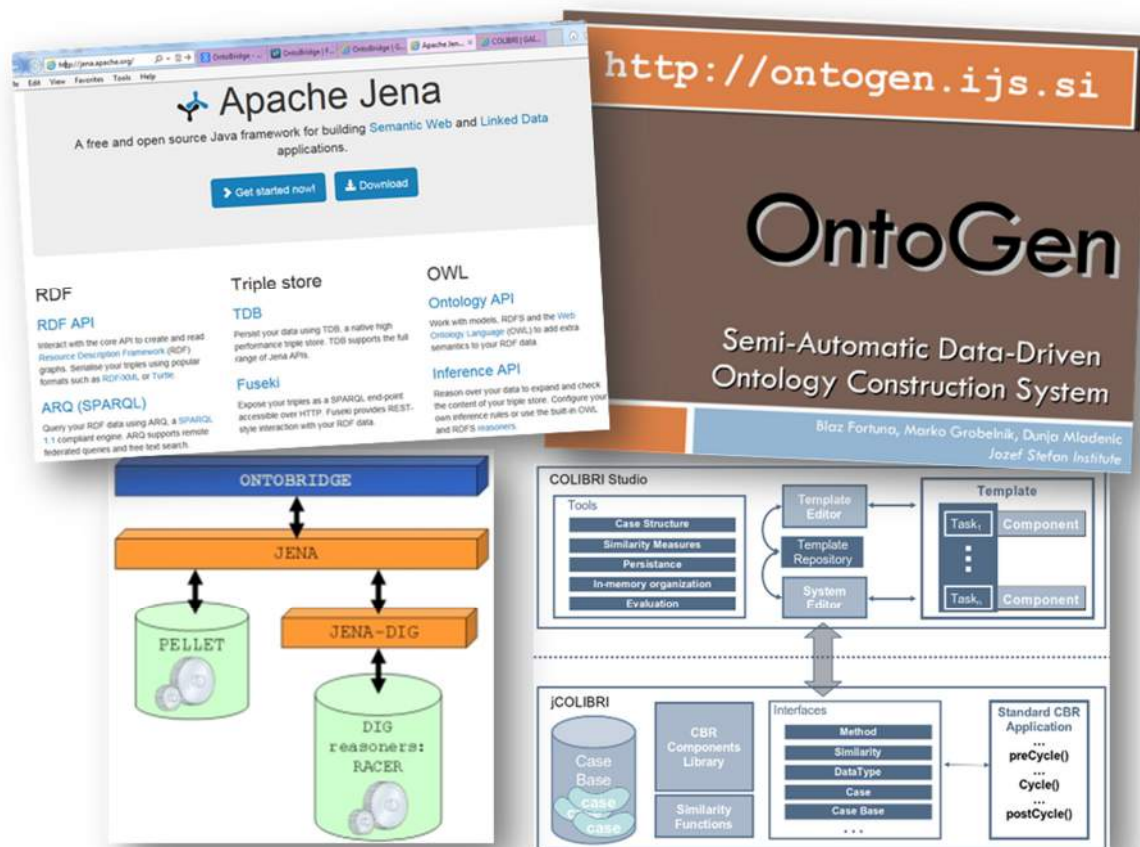


Figure 41: Jena, OntoGen, OntoBridge and COLIBRI

COLIBRI and jCOLIBRI⁵⁵

COLIBRI is a platform for building services that implement case-based reasoning (CBR) applications. Its developers in the Universidad Complutense de Madrid, Spain say that COLIBRI has semi-automatic configuration tools that make it possible to create a CBR system without writing a line of code. jCOLIBRI is a reference implementation of the COLIBRI architecture. OntoBridge, discussed later in this section of the deliverable is a subproject of jCOLIBRI.

Apache Jena⁵⁶

Apache Jena is a framework for building semantic web and linked data applications. It has API's to interact with RDF (resource description framework) graphs as well as OWL (Ontology Web Language) files. TIMBUS uses both of these constructs in its context model but does not employ Apache Jena. A nice feature of Jena is its support to query the ontology via ARQ (SPARQL) and Fuseki. ARQ is a query engine for Jena that supports the SPARQL RDF Query language, SPARQL is the query language developed by the W3C RDF Data Access Working Group and Fuseki is a SPARQL server which provides REST-style SPARQL HTTP Update, SPARQL Query, and SPARQL Update capabilities.

OntoBridge

OntoBridge is a java library that helps with ontology management. It uses text identifiers to manipulate ontology elements and is built on top of Apache Jena and it is a subproject of jCOLIBRI, both introduced earlier in this section of the deliverable. OntoBridge Supports:

Pellet⁵⁷ (internally) and DIG⁵⁸ complaint reasoners. Pellet is a java based OWL 2 reasoner and DIG (Description logic Implementation Group) is an interface which defines an XML schema which can be used to send messages between an application and a reasoned over an http connection.

OntoGen⁵⁹

OntoGen is an ontology editor which helps users by reducing the time spent and the complexity of generating ontologies by providing suggestions to the developer around concepts, concept relationships and visualisation to mention a few of its benefits.

5.2.3 Other Emerging Tools

DEX⁶⁰

DEX is an expert suite which assists decision making and the evaluation of different options developed by the Jožef Stefan Institute in Ljubljana. It carries out decision making by evaluating attribute values but one of the interesting parts of its approach is that it uses qualitative rather than quantitative attribute values to do this.

Enrycher⁶¹ **and** ***AnswerArt***⁶²

Enrycher and AnswerArt are computational linguistics tools developed by the Artificial Intelligence Laboratory in Slovenia. They process text to extract relations between topics and keywords and this can be done in a time series.

Figure 42, below, illustrates the output of Enrycher for that sentence. Relationships are identified between keywords and topics and these are grouped together. This sort of capability could be useful as an ingest tool in a TIMBUS-like repository if that repository contained documents relating to the operation and functioning of the process. These documents would essentially be items such as user guides and other collaterals which would assist future designated communities to use the re-deployed system. If these capabilities also extended cross-documents they could help the users discover related documents by identifying similar topics within them. It is often the case that different collaterals may exist for the different roles and functions of the preserved environment and a future user may require snippets of knowledge from a large cross-section of documents in order to perform a specific function.

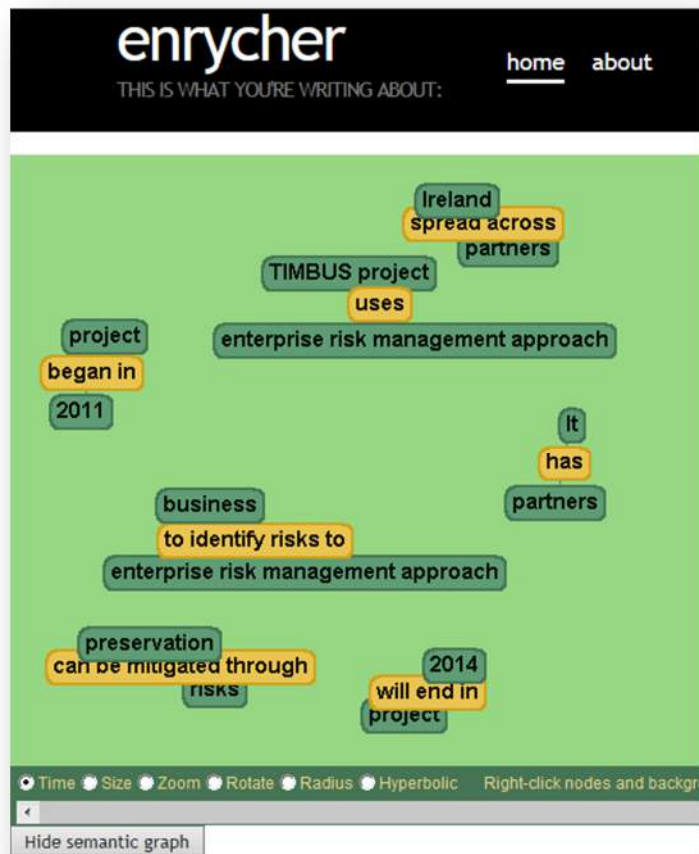


Figure 42: Output of Enrycher Web API

5.3 Conclusions

It is appropriate to consider big-data and other emerging tools and technologies given the reasonable expectation that any scalable solution will ultimately need to have the capability of preserving systems that store billions of files. These files all have metadata associated with them which needs to be indexed and searchable. The files themselves may be structured, semi-structured or unstructured data and they will contain everything from text files, to binaries, to source code, to images and videos.

Mitja Jermol argues in favour of an anti-corruption toolkit consisting of several of the technologies considered in this section as well as a few others which were not covered here. Many of the drivers for the TIMBUS approach to digital preservation are legalities based. They may be related to patent litigation, audit and/or regulatory requirements. This covers a huge range of scenarios and many of these could easily have the need to prove, in the eyes of the law, that a redeployed business process constitutes an accurate representation of the originally archived version in the repository and that the re-deployed process has not been subverted, or corrupted, in some manner to create a misrepresentation of the originally archived business process. This misrepresentation could be made more difficult by the analytical and metadata extraction capabilities of the tools reviewed in this section.

7 Conclusion

This document has presented the reader with an updated assessment of the digital preservation, highlighting both new and previously existing trends. Its intent was to familiarise anyone interested long term data retention on the current state of the art in this area. The report helps the reader by pulling together a collection of data available from other sources into one place and combining that with an informed assessment of trends being seen from the perspective of the TIMBUS FP7 Consortium.

The market assessment shows that the driving forces behind the need for digital preservation are increasing with little sign of this changing. The cost of storage, and hence archival solutions, is not increasing as rapidly as it was in last year's report, but it is still increasing. The larger diversity and perverseness of technology devices results in products with shorter lifecycles. The technologies which run on these devices are tied to point-in-time hardware and are either evolving or becoming extinct. In either case, they are changing more rapidly than ever before. It appears that it is not only technology change, but the rapid pace of that change which exacerbates the digital preservation challenge. Of equal importance is either the lack of awareness of the issue among decision makers, or the lack of budget to fund solutions. The latter can be considered a function of the former in many cases because the absence of appropriate awareness feeds in to the unwillingness to divert limited funds to an activity which is not properly understood.

Appendix A: Market-Watch Bibliography (Links to Referenced Material)

| Date | Summary | Source(s) |
|----------|--|---|
| 21/02/14 | Will Today's Digital Movies Exist in 100 Years? | http://spectrum.ieee.org/consumer-electronics/standards/will-to-days-digital-movies-exist-in-100-years |
| 13/02/14 | Print Books vs. E-books: What's the Future of Reading? | https://www.surveymonkey.com/blog/en/blog/13/03/25/print-books-vs-e-books-whats-the-future-of-reading/ |
| 12/02/14 | iRODS v4.0 is due for release in March 2014 | http://irods-consortium.org/dev/wp-content/uploads/13/02/irods-intro.pdf |
| 12/02/14 | The Storage Evolution: From Blocks, Files and Objects to Object Storage Systems | http://www.snia.org/sites/default/education/tutorials/2008/spring/storage/Bandulet-C_The_Storage_Evolution.pdf |
| 12/02/14 | WW1 soldier diaries published online - crowd sourcing their digitisation | http://www.bbc.co.uk/news/uk-25716569 |
| 06/02/14 | Trinity, Google Maps digitise Fagel map collection | http://www.techcentral.ie/trinity-google-maps-digitise-fagel-map-collection/ |
| 03/02/14 | Top 5 questions to ask before outsourcing your archives to a cloud vendor | http://trumantechologies.com/blog/ready-archive-cloud |
| 14/01/14 | Roger Highfield, Director of External Affairs at the Science Museum in London gave a short interview on George Hook recently talking about the topic of digital preservation so it shows just how main stream and ready for public consumption the topic is: | http://www.newstalk.ie/Keep-Your-Data-Safe |
| 09/01/14 | Imagine if you could surf Facebook ... from the Middle Ages. Well, it may not be as far off as it sounds. In a fun and interesting talk, researcher and engineer Frederic Kaplan shows off the Venice Time Machine, a project to digitize 80 kilometers of books to create an information system of Venetian history across 1000 years | Frederic Kaplan: How to build an information time machine |
| 08/01/14 | Government Notifies Standard For Digital Preservation Of e-Governance Data | http://www.siliconindia.com/news/technology/Government-Notifies-Standard-For-Digital-Preservation-Of-eGovernance-Data-nid-159159-cid-2.html |
| 07/01/14 | CES 2014: Sony shows off life logging app and kit | http://www.bbc.co.uk/news/technology-25633647 |
| 01/01/14 | Data Storage Innovation Conference | http://www.snia.org/events/dsicon2014?utm_source=SNIA+Email+List&utm_campaign=1ab799dab6-3rd_CFP_DSI_12_17_2013_copy_01_1_1_2014&utm_medium=email&utm_term=0_28326954a0-1ab799dab6-53360005 |
| 29/12/13 | BBC News: Classic 70s and 80s games go online | http://www.bbc.com/news/technology-25527786 |
| 21/12/13 | Laser Archaeology | http://ngm.nationalgeographic.com/13/12/laser-archaeology/johnson-text?utm_source=Twitter&utm_medium=Social&utm_content=link_tw20131221ngm-archtext&utm_campaign=Content |
| 17/12/13 | Impala: another google inspired platform | http://techcrunch.com/13/12/15/impala-another-google-inspired-platform-enters-the-mainstream-data-world/ |
| 16/12/13 | Nice Presentation by SNIA compares cross protocol storage standards (i.e. S3, Swift, CDMI, HDFS and Web-DAV) | http://snia.org/sites/default/files2/SDC2013/presentations/Cloud/ScottHoran_Lessons_Learned_Implementing.pdf |
| 13/12/13 | Ford set to digitize material from its archives for an online museum | http://www.freep.com/article/131212/BUSINESS/312120161/Ford-set-digitize-material-from-its-archives-an-online-museum |
| 09/12/13 | Announcing Suro: Backbone of Netflix's Data Pipeline | http://techblog.netflix.com/13/12/announcing-suro-backbone-of-netflixes.html?m=1 |
| 30/11/13 | Autodesk Launches New Tool for Digital Preservation ArchDaily | http://www.archdaily.com/452845/autodesk-launches-new-tool-for-digital-preservation/ |
| 30/11/13 | Browsers expiring over time | http://en.wikipedia.org/wiki/File:Usage_share_of_web_browsers_(Source_StatCounter).svg |
| 25/11/13 | Social media: The next generation of archiving | http://fcw.com/articles/13/11/25/executech-social-media-archiving.aspx |
| 24/10/13 | document from Mercedes showing how many open source licenses are actually used in the software that powers/drives/ships with modern cars | http://www4.mercedes-benz.com/manual-cars/ba/foss/content/en/assets/FOSS_licences.pdf |
| 21/10/13 | Back up that castle: Digital preservation group makes 3D copies of world's landmarks | http://www.startribune.com/nation/228616831.html |
| 21/10/13 | Million-Year Data Storage Disk Unveiled | http://www.technologyreview.com/view/520541/million-year-data-storage-disk-unveiled/ |
| | Same as above, just a different report | http://www.gizmag.com/billion-year-data-storage/29530/?utm_source=Gizmag+Subscribers&utm_campaign=a9a5ac8942-UA-2235360-4&utm_medium=email&utm_term=0_65b67362bd-a9a5ac8942-91281145 |
| 11/10/13 | CED emerging technology roadmap 2013-2016 | https://timbus.teco.edu/svn/timbus/work package 2/T2.3%20Exploitation%20plan/Documents/ |

| | | |
|----------|---|---|
| 27/09/13 | Perseids: a project belonging to the Parseus Digital Library. | http://sites.tufts.edu/perseids/ |
| 27/09/13 | P1484.13.2/D10, Sept 2013 - IEEE Approved Draft Recommended Practice for Learning Technology - Metadata Encoding and Transmission Standard (METS) Mapping to the Conceptual Model for Resource Aggregation | http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6603249 |
| 27/09/13 | MUSICAL PRESERVATION and Elvis's lost tapes | http://www.theatlantic.com/entertainment/archive/13/09/just-how-much-of-musical-history-has-been-lost-to-history/279948/ |
| 04/09/13 | Will Your Data Still Be Around Tomorrow? | http://www.forbes.com/sites/xerox/13/09/04/from-clay-tablets-to-electronic-tablets-preserving-content-and-knowledge-over-time/ |
| 02/09/13 | Add an expiration date to your tweets using a simple hashtag | http://www.theverge.com/13/9/2/4686228/time-your-tweets-to-disappear-with-using-a-simple-hashtag-twitterspirit |
| 27/08/13 | How Big Data is changing the world | http://www.bbc.co.uk/news/technology-23253949 |
| 19/07/13 | Amazon, MS cloud outages | http://www.bbc.co.uk/news/technology-23762526 |
| 16/07/13 | Good ol' tape will survive cloud era, too! | http://www.ciol.com/ciol/features/189797/tape-survive-cloud-era |
| 16/07/13 | History of magnetic tape | http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/tapestorage/ |
| 10/07/13 | They have some factors for sustainability here. | http://blogs.loc.gov/digitalpreservation/13/06/why-cant-you-just-build-it-and-leave-it-alone/ |
| 08/07/13 | London based company using ontologies for semantic search | http://www.ontology.com/ |
| 27/06/13 | Intel's taking a serious look at object storage. What's their game? | http://www.theregister.co.uk/13/06/27/intel_chipping_away_at_objects/ |
| 08/06/13 | Spar Point Group | http://www.sparpointgroup.com/ |
| 08/06/13 | CyArk is a 501c3 non profit organization with the mission of: digitally preserving cultural heritage sites through collecting, archiving and providing open access to data created by laser scanning, digital modeling, and other state-of-the-art technologies | http://archive.cyark.org/about |
| 03/06/13 | Tessella's digital preservation service available via G-Cloud | Tessella's digital preservation service available via G-Cloud |
| 30/05/13 | EDF Forum (European Data Forum) | http://13.data-forum.eu/program http://www.slideshare.net/EUDataForum |
| 30/05/13 | Duraspace demo | http://duraspace.org/tomorrow-529-duracloud-brown-bag-duracloud-glacier-1230pm-et |
| 27/05/13 | BBC scraps multi-million pound archive project | http://www.techweekeurope.co.uk/news/bbc-kills-digital-media-initiative-117244 |
| 06/05/13 | Spectra Logic Backup and Archive Blog | http://www.spectralogic.com/blog/index.cfm/13/4/16/Tape-NASThe-Storage-Game-Changer |
| 06/05/13 | A Measurement Framework for Evaluating Emulators for Digital Preservation | http://web.ebscohost.com/ehost/detail?sid=dc1243a6-76b8-452c-a8f1-c7b66c333188%40ses-sionmgr104&vid=1&hid=122&bdata=JkF1dGhUeXBIPWl-wLHVybCx1aWQmc2l0ZT1laG9zdC1saXZl#db=syh&AN=76258983 |
| 06/05/13 | A Service-Oriented Approach to Assess the Value of Digital Preservation by ENSURE project | http://rd.springer.com/chapter/10.1007/978-3-642-37804-1_17 |
| 06/05/13 | Springer series book from TIMBUS | http://rd.springer.com/article/10.1007/BF03323472 |
| 01/05/13 | CERN scientists launch project to restore world's first website - Irish Innovation News – Siliconrepublic.com | http://www.siliconrepublic.com/innovation/item/32480-cern-scientists-launch/ |
| 01/05/13 | views in the debate around the retention of personal data on the internet and how it could be data mined | http://www.libdemvoice.org/the-independent-view-data-preservation-instead-of-data-retention-34314.html |
| 27/04/13 | The Long Hill library in New Jersey ran a DP workshop as a fundraiser to pay for a new digital imaging service | http://newjerseyhills.com/echoes-sentinel/news/digital-preservation-workshop-at-long-hill-library-to-cover-photos/article_6c25e0f2-acf0-11e2-b1ec-001a4bcf887a.html?mode=jqm |
| 27/04/13 | Google afterlife | http://www.independent.co.uk/life-style/gadgets-and-tech/news/google-death-manager-search-engine-giant-lets-you-plan-digital-afterlife-with-inactive-account-manager-tool-8569848.html |
| 25/04/13 | RALEIGH, NC April 24, 2013 – The State Library of North Carolina, in conjunction with the State Archives of North Carolina, is releasing a redesigned, streamlined and mobile friendly digital preservation education site. | digitalpreservation.ncdcr.gov |
| 17/04/13 | Using tape as a NAS medium | http://www.spectralogic.com/blog/index.cfm/13/4/16/Tape-NASThe-Storage-Game-Changer |
| 17/04/13 | Microsoft secure Azure Storage goes down WORLD-WIDE | http://www.theregister.co.uk/13/02/22/azure_problem_that_should_never_happen_ever/ |
| 17/04/13 | Amazon margins on data storage | http://www.theregister.co.uk/13/04/02/amazon_drop-box_cloud_clone/ |
| 17/04/13 | DHSR's blog, Amazon margins on data storage | http://blog.dshr.org/13/04/more-on-amazons-margins.html |
| 10/04/13 | Cambridge capturing the digital universe | http://www.businessweekly.co.uk/hi-tech/15235-cambridge-captures-the-digital-universe |

Appendix B: Bibliography (Links to Referenced Material)

¹ ICT Telecommunications Development Bureau: <http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013->

[e.pdf](#)

² Internet Usage by Time Zone graphic sourced from <http://rohidassanap.wordpress.com/2013/02/14/how-many-internet-users-are-there-in-your-time-zone-infographic/> and is based on data gathered by www.pingfom.com.

³ Cisco VNI Global Internet Traffic: http://ciscovni.com/vni_forecast/advanced.html

⁴ Kleiner Perkins Caufield & Byers (KPCB) report on Internet Trends in 2013: <http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013>

⁵ Interview with BlackBerry CEO Thorsten Heins: <http://www.bloomberg.com/news/2013-04-30/blackberry-ceo-questions-future-of-tablets.html>

⁶ PEW Research Centre, Device Ownership statistics: [http://www.pewinternet.org/Trend-Data-\(Adults\)/Device-Ownership.aspx](http://www.pewinternet.org/Trend-Data-(Adults)/Device-Ownership.aspx)

⁷ SurveyMonkey Blog by Kayte Korwitts <https://www.surveymonkey.com/blog/en/blog/2013/03/25/print-books-vs-e-books-whats-the-future-of-reading/>

⁸ MyFitnessPal homepage: <http://www.myfitnesspal.com/>

⁹ Chris Woods of Intel Labs Europe speaks about Data as a Service for Social Change: <http://eitfoundation.org/innovation-forum/publication.html>

¹⁰ Social Media Top Sites as published by Ebzimba.com: <http://www.ebizmba.com/articles/social-networking-websites>

¹¹ BBC Story about WW1 Soldier Diaries going online: <http://www.bbc.co.uk/news/uk-25716569>

¹² United States Federal Records Act: <https://www2.ed.gov/policy/gen/leg/fra.html>

¹³ Worldwide Storage Software 2013–2017 Forecast and 2012 Vendor Shares: <http://www.idc.com/getdoc.jsp?containerId=243435>

¹⁴ FCW Article on the next Generation of Archiving: <http://fcw.com/articles/2013/11/25/exectech-social-media-archiving.aspx>

¹⁵ Maureen Pennocks Web Archiving Report: http://www.dpconline.org/component/docman/doc_download/865-dpctw13-01pdf

¹⁶ DPC DP Technology Watch Series: <http://www.dpconline.org/advice/technology-watch-reports>

¹⁷ DROID website: <http://digital-preservation.github.io/droid/>

¹⁸ PLATO preservation planning tool: <http://www.ifs.tuwien.ac.at/dp/plato/intro.html>

¹⁹ DSHR blog on URL shortners on the internet: <http://blog.dshr.org/2013/07/bitly-s-plan-z.html>

²⁰ Google Reader Service to End: http://bits.blogs.nytimes.com/2013/03/14/the-end-of-google-reader-sends-internet-into-an-uproar/?_php=true&_type=blogs&_r=0

²¹ Memento: <http://mementoweb.org/>

²² Architecture image sourced from: <http://www.mementoweb.org/guide/quick-intro/>

²³ Statistica: <http://www.statista.com/>

²⁴ Source for Mobile Web Usage: <http://rowanw.wordpress.com/2013/08/21/mobile-internet-usage-grows-36-year-on-year/>

²⁵ NodeJS homepage: <http://nodejs.org/>

²⁶ Sources for the four diagrams shown were: KPCB 2013 Internet Trends 2013, for browser usage graphic: <http://www.win-beta.org/news/internet-explorer-popular-north-america-google-chrome-still-dominant-browser-wars> and for programming languages graphic: <http://technologyfront.com/journalism/2012/2013-01/09.html>

²⁷ IDC Report on IoT 2013-2020 forecast: <http://www.idc.com/getdoc.jsp?containerId=243661>

²⁸ Graphic sourced from: <http://dupress.com/articles/rising-tide-exploring-pathways-to-growth-in-the-mobile-semiconductor-industry/>

²⁹ Tech vendors and cable companies push for more Wi-Fi spectrum: <http://arstechnica.com/information-technology/2014/02/tech-vendors-and-cable-companies-push-for-more-wi-fi-spectrum/>

³⁰ Jim Stogdill BIO: <http://radar.oreilly.com/jims>

³¹ Glen Martins Forbes Article on IoT: <http://www.forbes.com/sites/oreillymedia/2014/02/10/more-1876-than-1995/>

³² Worldwide Enterprise Storage Systems 2012–2016 Forecast (May 2012): <http://www.idc.com/getdoc.jsp?containerId=234990>

³³ Worldwide Enterprise Storage Systems 2012–2016 Forecast (May 2012): <http://www.idc.com/getdoc.jsp?containerId=234990>

³⁴ Worldwide Storage Software 2013–2017 Forecast and 2012 Vendor Shares: <http://www.idc.com/getdoc.jsp?containerId=243435>

³⁵ Worldwide Enterprise Storage Systems 2013–2017 Forecast: Customer Landscape Is Changing, Defining Demand for New Solutions: <http://www.idc.com/getdoc.jsp?containerId=241033>

³⁶ Dell Whitepaper: Object Storage

A Fresh Approach to Long-Term File Storage http://partnerdirect.dell.com/sites/channel/en-ca/documents/object_storage_overview_whitepaper.pdf

³⁷ IDC's Worldwide Cold Storage Ecosystem Taxonomy, 2014: <http://www.idc.com/getdoc.jsp?containerId=246732>

³⁸ Informatica Homepage: <http://www.informatica.com>

³⁹ Informatica Blogs on Data Archival: <http://blogs.informatica.com/perspectives/2013/03/06/enterprise-data-archiving-is-white-hot/> and <http://blogs.informatica.com/perspectives/2013/04/01/is-the-data-explosion-impacting-you-how-do-you-compare-to-your-peers/>.

⁴⁰ Michael Patterson's, SNIA Consultant, Article on preservation in the datacentre: http://www.ltdprm.com/Digital_Preservation_in_the_Datacentre.html

⁴¹ Michael Patterson's, SNIA Consultant Website article on using Cloud as an Archive solution: <http://www.ltdprm.org/reference-model/preservation-in-the-cloud/cloud-archive-requirements>

⁴² Article from the Register about Nirvanix going out of business: http://www.theregister.co.uk/2013/09/19/cloudy_storage_startup_nirvanix_sinking/

⁴³ End-of-life in the cloud: <http://blog.dshr.org/2013/10/end-of-life-in-cloud.html>

-
- ⁴⁴ Truman Technologies Blog Article on Outsourcing archives to a cloud vendor: <http://trumantechnologies.com/blog/ready-archive-cloud>
- ⁴⁵ David Rosenthal Blog on DP costs: <http://blog.dshr.org/2013/07/talk-at-digital-preservation-2013.html>
- ⁴⁶ Chris Taylors wired.com article about the 3 V's of Big Data: <http://www.wired.com/insights/2013/08/three-enormous-problems-big-data-tech-solves/>
- ⁴⁷ Big Data Analytics Begins with Intel: <http://www.intel.com/content/www/us/en/big-data/big-data-analytics-turning-big-data-into-intelligence.html?cid=sem116p40188g-c&gclid=CKeF5fTDnr0CFS7MtAodD1MA-g>
- ⁴⁸ HANA. Not just a database but a whole new approach to data: http://global.sap.com/campaign/na/usa/CRM-US12-PPC-PPCANAC9/index_ifx.html?SOURCEID=DE&campaigncode=CRM-XU14-INT-PPCNMEM1DE&utm_source=google&utm_medium=ppc&utm_term=big%2520data&utm_campaign=In-Memory%2520A%252FC&mid=smOrjUtSM%7cdc_2722p1v19626_34845801537_big%2520data_p&kwid=mOrjUtSM
- ⁴⁹ The Intel-HANA Story: <http://hadoop.intel.com/videos/idh-sap-hana-story>
- ⁵⁰ Netflix Suro: <http://techblog.netflix.com/2013/12/announcing-suro-backbone-of-netflix.html?m=1>
- ⁵¹ Cloudera Impala: <http://techblog.netflix.com/2013/12/announcing-suro-backbone-of-netflix.html?m=1>
- ⁵² Mitja Jermol's EPAC presentation on Anti-corruption: http://www.epac.at/downloads/info-materials/doc_download/74-anti-corruption-workshop-mitja-jermol-presentation
- ⁵³ CycCorp homepage: <http://www.cyc.com/>
- ⁵⁴ LarkC FP7 project: <http://www.larkc.eu>
- ⁵⁵ COLIBRI: <http://gaia.fdi.ucm.es/research/colibri>
- ⁵⁶ Apache Jena: <http://jena.apache.org/>
- ⁵⁷ Pellet: <http://clarkparsia.com/pellet/>
- ⁵⁸⁵⁸ DIG 2.0: <http://dig.cs.manchester.ac.uk/overview.html>
- ⁵⁹ OntoGen: <http://ontogen.ijs.si/>
- ⁶⁰ DEX: <http://www.ai.ijs.si/MarkoBohanec/dex.html>
- ⁶¹ Enrycher: <http://ailab.ijs.si/tools/enrycher/>
- ⁶² AnswerArt: <http://ailab.ijs.si/tools/answerart/>