talis

Unleashing the true value of content

PROJECT Bigfoot

A white paper

by Justin Leavesley
Chief Technical Officer, Talis

	Overview	4
1.	Unleashing the true value of content	5
2.	The theory of sharing content	6
2.1	The limitations of content today	6
2.2	The possibilities for content sharing in the future	6
2.3	The theory of sharing content: Web 2.0	6
2.3.1	The Amazon website experience	6
2.3.2	The rise of Web 2.0	6
2.3.3	How using Web 2.0 helps create richer content	7
2.3.4	The challenges presented by Web 2.0	7
2.3.4.1	The challenge of exporting data and functions	7
2.3.4.2	The challenge of seperating data and functions from specific systems	7
2.3.4.3	The challenge of legacy systems	7
3.	The reality of sharing content: Project Bigfoot	8
3.1	The dynamics of content delivery	8
3.1.1	Content providers need to provide content	8
3.1.2	Content consumers need content to be available transparently	8
3.2	Bigfoot is the enabler of content sharing	8
3.3	How it works	9
3.3.1	Functional decoupling of consumer and provider	9
3.3.1.1	Sharing by copying	9
3.3.1.2	Sharing by reference	9
3.3.1.3	Functional decoupling using Silkworm foundation services	9
3.3.2	Non-functional decoupling using Bigfoot	9
3.3.2.1	Flexible approach to data distribution	10
3.3.2.2	Hybrid query processing	11
4.	What are the benefits of Bigfoot?	11
4.1	Bigfoot: Unleashing the true value of content	11
4.2	Bigfoot: A bridge between systems	11
4.3	So what exactly could Bigfoot enable?	11
4.3.1	Sharing and participation in content collections	11
4.3.1.1	Using Web 1.0	12
4.3.1.2	Using Web 2.0 and Bigfoot	12
4.3.2	The power of Aggregation	12
4.3.2.1	Using Web 1.0	12
4.3.2.2	Using Web 2.0 and Bigfoot	12
4.4	New economics of resource discovery	13
4.4.1	Traditional resource discovery	13
4.4.2	A new type of resource discovery	13
4.4.3	Available through the Bigfoot and Silkworm platform	13
4.4.4	More cost-effective resource sharing	13
4.4.5	More scaleable resource sharing	13
4.4.6	Access to more widely held resources	13
5.	Conclusion	14
	About Talis	15
	About the Author	15

Overview

A world where richer content is simply available

Consider this scenario. A library's collection contains a rare and precious book. In theory, the content of the book is truly valuable. In reality, a search in Google doesn't produce a single result for it. In fact, very few people know that the book is available.

Potentially the book may have value but, for most people, it might as well not exist.
Sadly, this situation is all too common. And it's not restricted to books. Historically, many forms of content, whether that's metadata, videos, images or text, are trapped within islands of content. Different systems hold different content in different ways.

Now imagine an alternative where multiple applications all made use of the same core data, reusing and remixing it to serve the user's needs – whatever those were.

The result would be that data could be used far more dynamically and become part of a richer, better experience for its consumers.

Talis Bigfoot, using Web 2.0 technologies, offers a way of removing these silos and sharing content more freely by creating an intermediation layer. This platform hides the technical complexities of how content can be shared among multiple applications – making it easier for content providers to deliver content and enhancing how content consumers receive it.

Thanks to Bigfoot, every content provider won't need to create its own version of this platform. And content consumers won't need to understand complex technologies to gain access to rich content.

1: Unleashing the true value of content

Today, content is king. The huge success of Internet services such as Google, Amazon and Yahoo! show that users are now demanding experiences that are based on rich content. The expectations of this 'Google Generation' are geared towards an experience that is:

- All Encompassing
- Without Walls
- Immediate
- Rich
- Participative
- Personalised
- Fulfilling

Content, whether it's metadata, video, images or text documents, therefore has the potential to be a hugely valuable asset. However, the true value of data may have very little to do with the quality of the data itself. Instead, much depends on how that data is managed and how easy it is to access.

Quite simply, however valuable data is, if it isn't available, it might as well not exist.

Consider, for example, the fact that, more often than not, data and content is only available from one or a small number of specially built applications (including web applications). Because of that, its value is limited only to what these applications can do.

What if there was another way? What if multiple applications could make use of the same data, reusing and remixing it for new purposes? Imagine the possibilities if multiple applications could orchestrate content from multiple systems in order to serve the needs of the user – whatever those were. If data could be used wherever and whenever there was a point in doing so, including being part of an orchestration of content to create new, synthesised, experiences, then content would become truly valuable.

2: The theory of sharing content

2.1 The limitations of content today

Currently, though, content cannot be accessed in this dynamic way. Instead data is still fragmented across islands of content that the user has to visit in turn. With this content confined to silos, its value is confined to the limits of each system.

The rise of the web has made it easier than ever to access information systems, but the first generation of the web (Web 1.0) does not actually release data from the confines of the silo, it merely makes it easier for a human to find and use the system.

2.2 The possibilities for content sharing in the future

Breaking down the walls between systems to create the applications that users are coming to expect requires collaboration between systems and systems vendors at a fundamental level. In the future, somehow, systems need to be able to share content.

2.3 The theory of sharing content: Web 2.0

2.3.1 The Rise of Web 2.0

This is possible because Web 1.0 is now evolving into Web 2.0, a network that supports machines as well as human interfaces. Using Web 2.0 means that content or transactions can be reused and built upon instead of being locked up in silos. Applications can literally reuse the functionality of other systems, remixing capabilities to build a new application.

Web 2.0 is a fundamental change in the technology stack that we have been living with for the last 15-20 years. This change is occurring due to universal connectivity and standardisation combined with the massive increase performance/price.

In the early days of the Internet, it was simply bolted onto our existing technology stack. The Internet connected applications together. From the application point of view, the Internet was on the outside.

So more and more, the value of the application is in the assembly of web services. Moving the value of collaborating into the sharable standards based web services is the perfect way for vendors to build applications yet share and enrich a common value, breaking through the silos.

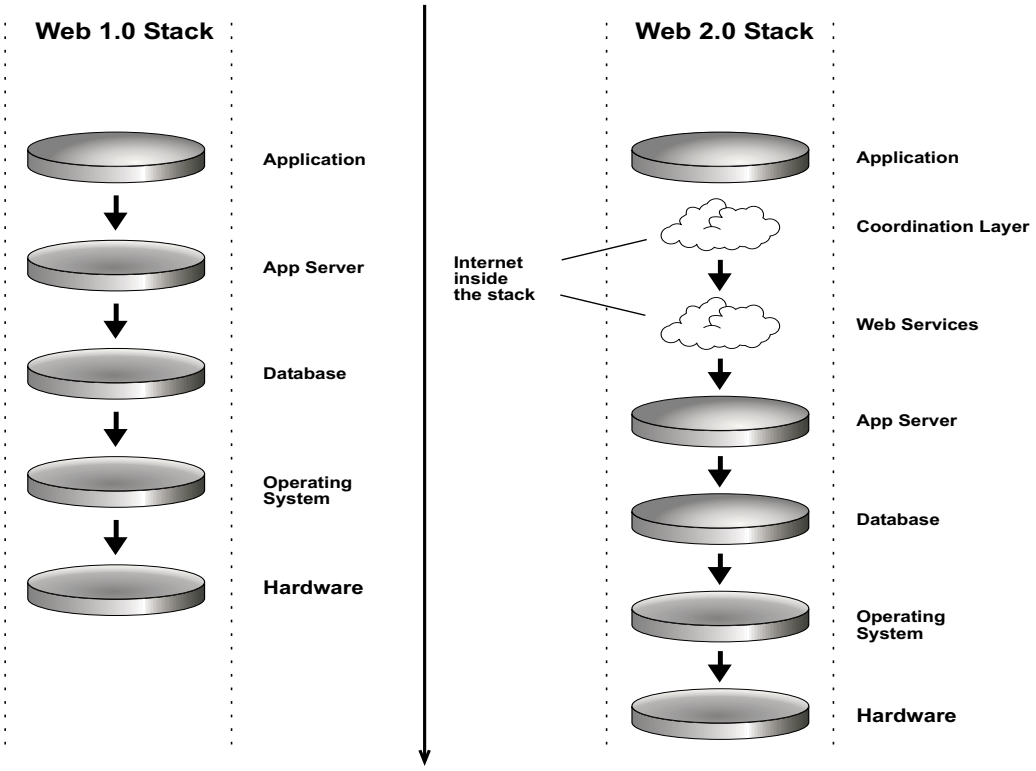
The reasons behind the rise of Web 2.0 are explored further in Talis' white paper on Project Silkworm.

2.3.2 How using Web 2.0 supports the sharing of content

Web 2.0 technologies massively reduce the barrier to sharing. Using Web 2.0 technology, content can be delivered to where it is needed within any application – immediately enabling traditional boundaries to content sharing to be removed.

2.3.3 How using Web 2.0 helps create richer content

That's not all either. Content can be combined with related data to create richer data and it can also be grouped, aggregated and mined to produce new derived information from the patterns and relationships contained within.



What's more, for systems that allow contribution from users (i.e. have an architecture of participation, such as the way Amazon includes buying history and invites reader reviews), this also fuels a massive network effect. As users get involved and contribute content to the system, this builds value for other users. The users themselves therefore enrich the content.

2.3.4 The challenges presented by Web 2.0

2.3.4.1 The challenge of exposing data and functions

There are the obvious functional issues of exposing data and functions via web services protocols. This is a particular challenge for older client server architectures where much of the application logic is locked in thick clients. If a system cannot be decomposed into a series of web components or at least one big web component then it cannot participate in the next generation of the web.

2.3.4.2 The challenge of separating data and functions from specific systems

Apart from the functional issues for legacy systems there are also significant non-functional issues preventing participation in Web 2.0. Consider the fact that most existing systems were built with quite specific user scenarios and bounded user load profiles.
This was because the clients of the data and functions were tightly bound as part of the system and therefore well understood. If, though, the data and functions of a system are suddenly made available as part of other applications, then completely new system loads can easily be generated that the original system was never designed to manage.

2.3.4.3 The challenge of legacy systems

Of course green field development can design with Web 2.0 in mind, but the vast number of existing systems cannot just be replaced overnight. Instead, ways must be found to allow them to become available as Web 2.0 components, in much the same way that legacy systems were given web interfaces as Web 1.0 became mainstream.

3. The reality of sharing content: Project Bigfoot

As with any technology, the theory only goes so far. Then, the reality must be considered. No matter how brilliant a technology, to succeed it must also be workable. The personal computer owes much of its success to the humble mouse – an innovation that made it user-friendly. In a similar way, translating Web 2.0 into a workable solution is the key challenge.

3.1 The dynamics of content delivery

Currently, content is delivered from content providers to content consumers. For the content sharing world that Web 2.0 enables to be successful, the exchange of information between the content providers and the content consumers must now be delivered in a completely different way.

3.1.1 Content providers need to provide content

Currently, content providers are responsible for the data in content collections. But, does that mean that each and every content provider has to invest in the technology, hardware and skills required to make their content available to every demand? If this were the case, it would represent a huge barrier for creating this new content-sharing world. For this idea to work, content providers need to be free to provide content – not to provide the sophisticated content-sharing technology

3.1.2 Content consumers need content to be available transparently

What's more, the reality for content consumers must be carefully considered. They need to be able to deal with logical collections of content without needing to worry about how the underlying systems are accessed, where they are located, or what load they are designed to cope with. Again, content consumers must be able to understand the content – not the technology behind it.

3.2 Bigfoot is the enabler of content sharing

Bigfoot is part of a network platform based on Web 2.0 that makes it easier to share content across multiple applications. It is a shared metadata and content platform that sits on top of the Silkworm foundation services and allows the simple creation and management of Web 2.0 data components. It also enables the grouping of content collection into aggregations that are themselves exposed as Web 2.0 data components.

At its most fundamental, the Bigfoot and Silkworm platform decouples the content consumer from the content providers by introducing an intermediation layer. This intermediation layer (the platform) hides the functional and non-functional complexities of the diverse landscape of existing content stores from the content consumer. This allows the content consumer to deal with data and content in a uniform way – without needing to understand the technology.
At the same time, thanks to the existence of Bigfoot, and the fact that it can work with multiple content stores, content providers can concentrate on what they do best – providing content.
By decoupling content consumers from content providers, Bigfoot and Silkworm reduces the barriers to participation for each by addressing both functional and non-functional issues associated with consuming and exposing data as web 2.0 components.

3.3 How it works

3.3.1 Functional decoupling of consumer and provider

3.3.1.1 Sharing by copying

In the past, if content was to be shared it would need to be copied into the original web page. Code would need to be compiled into an application or distributed with it. This causes many problems. Once copied, content is bulky and out-of-date and content providers may not know how to maintain the content so the quality may also be poor.

3.3.1.2 Sharing by reference

In contrast, sharing by reference enables a web page to contain references to other pages with live links to the page. It's the difference between emailing a link to an interesting web page and emailing the actual web page html.

Being able to share by reference means that the original content provider remains responsible for the continued maintenance and currency of the content. Sharing by reference is far more stable than sharing by copy since it is usual that the client copying the content does not know how to maintain the content. It also simplifies issues of data quality and ownership.

3.3.1.3 Functional decoupling using Silkworm foundation services

The Silkworm foundation services provide functional decoupling of consumer and provider through uniform data addressing and retrieval (UDAR) URL. Every record in a content node registered in the Silkworm platform has a unique, persistent URL that turns each record into a first class web resource. The UDAR URL is a live reference to that data record. If pasted into a browser (i.e. HTTP GET performed) the Silkworm directory, access and transformation services retrieve the record from the content system and return it as XML or other formats if requested (e.g. RDF). This allows systems to share the data by passing round live references to it rather than copying the actual content.

The UDAR URL coupled with the Silkworm foundation services allows data formally locked deep in hidden databases to be rescued by reference. Consider, for example, if a UDAR URL points to the XML describing the discography of a musician held in a music database. If the musician releases another CD then, as soon as the database is updated (perhaps by the publisher), all the systems using the reference to the discography will automatically also display the new record. The responsibility for content quality remains with the authoritative experts.

3.3.2 Non-functional decoupling using Bigfoot

However, it is easy to see how, if every reference results in a call to the original content providers system (the content node) then this model increases load and therefore required infrastructure of the content provider significantly. Issues of latency and availability then become important.

Bigfoot can help to avoid this problem. Today, we are used to the Internet allowing one computer to contact another without having to know or care about where the routers are, which switches are used and how the ISP has positioned web caches. We expect that the infrastructure of the network can be optimised without affecting the functional use of it.

Bigfoot also allows physical distribution of data, indexes and the processing of a query to be optimised without affecting the application layer. It simply doesn't need to know about it.

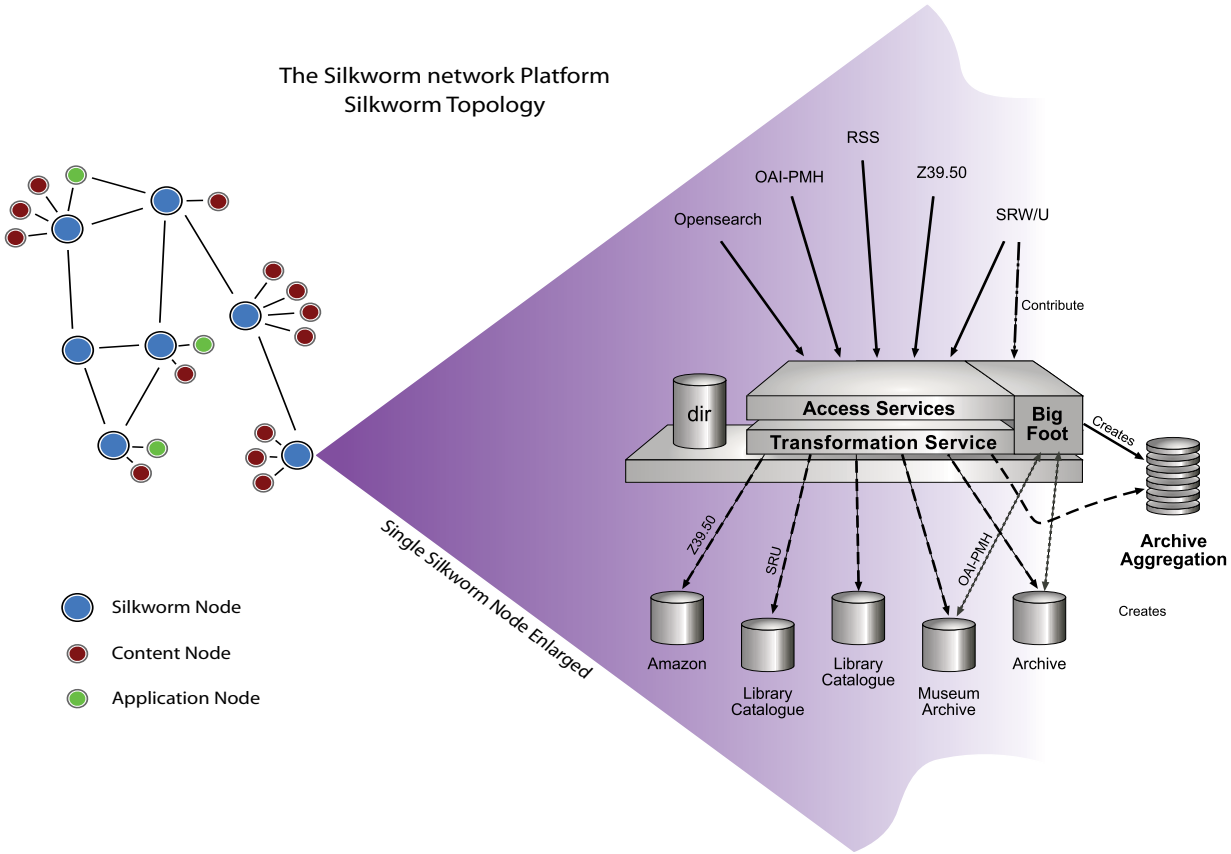
3.3.2.1 Flexible approach to data distribution

Bigfoot provides flexible approaches to data distribution. For example, if a content node supports harvesting, then Bigfoot can move the content into the Bigfoot cloud and direct searches to the harvested data in Bigfoot rather than the content node. Bigfoot maintains the currency of the data cache by frequent harvesting of only the changes made by the content node. In this way Bigfoot can remove the search and retrieval load from the content provider leaving the content providers' systems to focus on the quality of the content.

Bigfoot could distribute caches of referenced data to the closest Silkworm node to the client application. This could be a public node, maybe within the institution running the application, or could even be in the client application if the Silkworm node modules are compiled into the application (i.e. the application becomes a Silkworm node itself).

Through a mechanism of notification and deltas (changes) the local caches are kept in near real time step with the content provider.

Removing the cost of providing such a searching infrastructure allows niche content providers to make data available. This could promote a long tail of content provision to compliment the smaller number of high value content systems (which are likely to provide their own infrastructures).



3.3.2.2 Hybrid Query processing

An aggregation collection is one in which a search draws results from numerous content nodes into one. Bigfoot allows new logical collections to be defined, which are the aggregation of many other collections (including other aggregate collections).

Bigfoot hides the details of how the query is actually executed. For example, setting up an aggregation of library holdings may cause Bigfoot to harvest data from library systems that support harvesting protocols into a central store and combine this at query execution time with results from federated searching of library systems that only support search protocols. The aggregate collection appears as just another collection. The details of query processing are hidden and can be optimised as required with out affecting the client applications.

Similarly, rules for de-duplication and identity can be defined to ensure the aggregation operates are required.

4. What are the benefits of Bigfoot?

Thanks to Bigfoot, the barrier to both content consuming applications and content providers is lowered. This enables more participation and sharing, creating a network effect which increases the total value of applications and content for those involved.

4.1 Bigfoot: Unleashing the true value of content

Using Bigfoot, content can be made available so that it can be used outside of the current silo that is limiting its value.

4.2 Bigfoot: A bridge between systems

Bigfoot and Silkworm unload complexity from the application developer by encapsulating it in the platform. All the application developer needs to understand is the simple platform interfaces.

Bigfoot and Silkworm allows the application developer to work with logical collections of data that can be searched, referenced to and combined as required. The physical distribution of data, the location of content collections and the protocols required to access them are all hidden from the application layer.

An application can work with a logical aggregation of hundreds of separate collections. The details of how this is achieved, through a hybrid of federated search and physical union techniques is the concern of Bigfoot not the application developer.

In this way, Bigfoot massively reduces the barriers, cost and complexity of sharing data and provides a layer that bridges the differences between systems. Content consumers and content providers can concentrate on their own competencies – without needing to understand or create this technology themselves.

4.3. So what exactly could Bigfoot enable?

4.3.1 Sharing and participation in content collections

An organisation has collected reader book reviews over a number of years. This information is valuable to the Author, publishers and other readers trying to decide on the next book to read. The collection process was expensive involving direct interviews with the reader.

4.3.1.1 Using Web 1.0

With the advent of Web 1.0, the organisation had the opportunity to make its content more easily available to authors, publishers and readers with a subscription-based website. On the website, readers were invited to submit reviews and be involved without needing to subscribe.
However, although the website was very well used by the authors and publishers, the public was not keen to sign up and few reviews were posted.
The value of the reviews is therefore limited to the high end users of that content, the publishers and the authors through the human readable website. The value of the site may be less to readers than to publishers and authors but, inevitably, there are potentially hundreds of thousands more readers than authors or publishers. With Web 1.0, content owners are unable to exploit the potential of these people – the long tail of the review market – to enrich the content.

4.3.1.2 Using Web 2.0 and Bigfoot

If, however, the content was Bigfoot-enabled it could be embedded, live, into any application, be that a website or client application, that dealt with books. Then, the reviews would be available to the user and, consequently, of value to them. The submit review functionality could be available as a web service and embedded into any application that used the review content. Users would be able to submit reviews much more easily, with the result that the content would be enhanced.

By placing the third party applications in front of users (rather than making them seek it out in a separate silo), the reach of the content is extended. The long tail becomes economically harvestable through subscriptions for application developer or other means. And, of course, the more users involved, the more content is created and the greater the value of the collection as a whole. Plus, the content is shared live, so if a new review is added to a book in the collection, then it appears in the applications using that content collection.

4.3.2 The power of Aggregation

Consider a network of library management systems. Each system has a content collection of bibliographic descriptions for the works in stock.

4.3.2.1 Using Web 1.0

Individually each content collection is not extensive. The typical library collection only covers a small part of that data universe (i.e. all the published works). This results in the user having to search other systems to ensure the work is found or, indeed, that it definitely does not exist.

Aggregating the content from each of these collections into a single collection increases the coverage but the Web 1.0 technologies that do this are expensive and the resultant collection is not easily available to applications. (They may, for example, use Z39.50).

4.3.2.2 Using Web 2.0 and Bigfoot

Through Bigfoot and the Silkworm platform, a group of collections can be treated as a single collection by creating a Bigfoot aggregation node.
Bigfoot uses the directory, access services and transformation services of the Silkworm foundation platform to create an aggregation by combining harvesting and federated search across the individual collections. This aggregation is a Bigfoot Web 2.0 content component that can be easily used, via low barrier APIs, by any application.

Bigfoot is a platform for creating Web 2.0 content components, either via individual existing content nodes or through aggregating across many nodes. Because the complexity is managed by the platform, the cost and technical barriers for the applications built on Bigfoot are hugely reduced.

4.4 New economics of resource discovery

4.4.1 Traditional resource discovery

Talis Base and UnityWeb are union databases for bibliographic descriptions and library holdings respectively. They are used to share the effort of cataloguing amongst Talis Libraries and to discover the location of resources for InterLibrary Lending in the UK. Although robust, these systems were expensive to build and the approach to aggregation is costly to run. This cost structure was the result of the technology landscape ten years ago.

4.4.2 A new type of resource discovery

Project SkyWalk is a Talis research project that takes a fresh look at resource sharing technology. Building on the Silkworm and Bigfoot platform, SkyWalk will take full advantage of the unfolding Web 2.0 technology stack to enable new business models for resource sharing. The fundamental drivers of change are the radically lowered cost of hardware to host the service and the dramatic drop in complexity to build and run the system – thanks to the Bigfoot and Silkworm platform.

4.4.3 Available through the Bigfoot and Silkworm platform

The Silkworm directory will contain entries for libraries that are participating along with the technical details of protocols supported. Bigfoot will then use the Silkworm access services for uniform access to these disparate protocols in order to create and manage a Web 2.0 data component for bibliographic descriptions and holdings. These aggregations can then be searched or referenced just like any other Silkworm content collection, using a variety of protocols including ultra low barrier protocols such as REST.

For libraries that can support the harvesting of Deltas (incremental change) through protocols such as OAI-PMH, Z39.50 using the 005 marc tag, or simple export differencing software, Bigfoot will create a physical union database. This harvested database will be combined with federated search results for libraries that are not supporting any harvesting protocols.

4.4.4 More cost-effective resource sharing

This approach reduces costs so greatly that it becomes economic to consider making the resultant database content available under a creative commons license, and so free for non-commercial use.

4.4.5 More scaleable resource sharing

The resulting system is also inherently more scalable because it's built on a network platform of potentially many nodes, and interoperability with the existing union system is possible using the hybrid harvesting and federated search protocols as required. It would still be possible to receive submissions via the older methods of tape dumps but this manual approach has extra costs that would have to be covered.

4.4.6 Access to more widely held resources

Holdings from other non-traditional holdings systems, such as Amazon Marketplace, could be included in the SkyWalk holdings union. You could consider Amazon market place to be a circulation system where the stock is on shelves in second hand books stores and bedrooms.

A brief analysis of ILL requests in the UK public library sector revealed around 70% of requests could be serviced via the Amazon market place. Astonishingly, around 10% could be purchased for free (or, more accurately, just the cost of postage and packaging). This could allow some ILL requests to be fulfilled directly from Amazon at a fraction of the cost of traditional ILL processing and with a next day service potentially directly to the borrower's door. Once the borrower has finished with the book the library has the option to add the work to its stock or dispose of it, perhaps even through the Amazon market place again.

Because the bibliographic and holdings unions are Web 2.0 components, they can be easily included in third-party applications such as non-Talis ILS systems, ILL request management systems or web applications. Talis may create web applications to allow simple and direct user access to holdings and bibliographic records.

5. Conclusion

Moore's law rewrites the technical constraints governing software systems every five to ten years so that the complex and expensive becomes mundane. As the new technical landscape becomes clear, structures emerge to manage the complexity issues that arise. These form new layers in the software stack and give rise to platforms.

In this way, the Bigfoot platform is now enabling application developers to deal with content systems at a higher or logical level of abstraction. The complexities of data distribution, caching, location and a multitude of protocols are hidden from both the application developers and the content providers.

Consequently, optimisation can occur separately from the application layer. This is as essential for the success of Web 2.0 data availability, as it was for the Internet's communication optimisation to be separated from the application developer.

Talis will be using this platform to power its next generation of content applications as well as making this infrastructure available for other organisations to create their own applications.

About Talis

Talis is the leading provider of library management solutions for the UK and Ireland. Designed for the academic and public sectors, Talis' suite of products helps public authorities, universities and colleges to manage their extensive resources and provides users with swift and easy access to that information. Thanks to Talis' support for greater integration, libraries in both the public and academic sectors are also able to streamline multiple systems and improve internal efficiency.

Founded in 1969, Talis has grown steadily and now enjoys an impressive 25% share of the UK library market. The company also has a long history as a technology innovator and invests heavily in research into the information management technologies of the future. Ongoing research projects include metadata, web services, RSS and service-oriented architectures.

For more information, please visit www.talis.com

About the Author

Justin Leavesley - Chief Technology Officer

Justin joined Talis in January 2000 from Morse Group plc, a leading European technology supplier. After receiving his Master of Physics from Oxford University, Justin joined IBM working first in the defence and aerospace sector for the British Army and later for Natwest Bank.

Justin left IBM to join an internet start up company based at Birmingham University where he was involved in product development, marketing and business strategy. Justin, a Sun certified Java enterprise architect, is now responsible for technology and related business strategy at Talis where he leads the Research Group.