講演記録

# “The ISI Web of Knowledge℠ Platform: Current and Future Directions for Database Access and Linking”

Jeff Clovis（Director, International Product Support, ISI - Thomson）
（日本語サマリー作成　渡辺麻子　ISI－Thomson）

以下は、2001年９月に関西大学で行われた「Web of Science導入記念セミナー」での、Jeff Clovisによる講演の日本語サマリーである。Jeff ClovisはISI－Thomsonの製品サポート部門のディレクターで、長年データ作成や製品開発に携わってきた。講演では情報プラットフォームISI Web of Knowledgeについて発表したが、詳細については下記論文も参照されたい。http://www.isinet.com/isi/hot/essays/isiplatform/8105138/index.html

1955年Dr. GarfieldはCitation IndexesをScience誌上に発表した。Citation Indexesは引用による文献間のリンクに注目したもので、キーワードや主題分類による検索を超える新しい方法であった。すなわち引用による文献間のリンクをたどってゆくと、効率よく必要な文献が検索できるというものである。このアイディアを最初に実現したのはScience Citation Index®（以下SCI®）であった。SCIは様々なメディアで提供されてきたが、1997年にWeb of Science®（Citation IndexesのWeb版）が出現して初めて、研究者は引用文献検索の醍醐味を味わうこととなった。Web技術によって文献間のリンクが簡単にたどれるようになったのである。Web技術はその後の現在にいたるまで進化を遂げ、ISIのWeb製品を統合した情報プラットフォームWeb of Knowledge℠ に発展してゆくのである。

ISI Web of Knowledgeとは、ISIが過去40年間かけて築き上げてきたコンテンツとツールを抱合する、引用リンクを基盤とした情報プラットフォームである。研究者は自分のデスクトップから統一環境下で、ISIが厳選した高品質の情報にアクセスできる。このWeb of Knowledgeを支えているのは、MuscatDiscovery™とISI Linksという二つの技術である。

MuscatDiscoveryは、これまでの論理演算による検索と違って、検索者が欲しいとおもっている情報を推測し検索してくる。情報検索はMuscatDiscoveryによって大きく拡張された。データベースのような体系的データと、Web上のコンテンツのような非体系的データの両方が同時に検索できる。また、Plain Text、HTML、PostScript、PDFなどデータのタイプを問わずにまとめて検索できる。使い勝手も易しく、エキスパートでなくても検索できるのも大きな利点である。ISI Web of Knowledge上でMuscatDiscoveryを応用しているのは、ISI eSearchとISI CrossSearchである。ISI eSearchではWebサイトから厳選されたWebドキュメントのフルテキストを、文献情報検索と統合した方法で検索することができる。ISI CrossSearchは異なったタイプのコンテンツ（雑誌論文、会議録、特許、など）を単一のインターフェースで横断検索するものである。複数の情報源を単一の入力ポイントからシングルセッションで管理できる。

ISI Linksは、Web of Knowledge上のあらゆるタイプのリンクを管理する機能である。リンクには、intra-content links（引用文献からフルレコードへのリンクなど）、Inter content links（論文と特許のリンク、論文とDNA配列データのリンクなど）、電子ジャーナルリンク、OPACリンクなどがあるが、こういったものがすべて管理できる。電子ジャーナルリンクについて特に述べると、今までISIは出版社からリンク情報に基づいて、レコード単位で電子ジャーナルへのリンクをはっていた。出版社からのリンク情報によって電子ジャーナルリンクは安定したが、時間と労力を要した。一方、Algorithmic Linkingは、出版社による情報提供に基づくリンクとは異なったタイプの技術である。メタデータにもとづきリンク先の情報を推測する技術で、その結果クリックした瞬間に電子ジャーナルへのリンクが実現するのである。Algorithmic Linkingの問題点はリンクが完全に保証されないことだが、ISI はRoboLinksという技術によって、このAlgorithmic Linkingを信頼性とともに提供することができるようになった。システムの一部にリンクが生きているかどうか自動的に確認する機能がつく。Robolinksによってリンク情報を

提供できない出版社の電子ジャーナルもリンクすることができるようになるなど、リンクの可能性が大きく広がった。

ISI Web of Knowledgeはこのような技術によって、実現した情報プラットフォームである。この情報プラットフォーム上の個々の情報源も今後強化されてくる。Web of Scienceのバージョンアップ、BIOSIS、CABI、IEEの新しい情報源の搭載などが予定されている。ISI Web of Knowledgeは真のポータル環境を目指して発展し続け、データベースアクセスとリンクがともに進化してゆくことになるであろう。

# “The ISI Web of Knowledge℠ Platform: Current and Future Directions for Database Access and Linking”

The following text is based on a presentation at Kansai University, September 2001 by **Jeff Clovis, Director, International Product Support, ISI, 3501 Market Street, Philadelphia, PA 19104.**
A complete version of this paper is published on the ISI Web site at
http://www.isinet.com/isi/hot/essays/isiplatform/8105138/index.html.

In 1955, Dr. Garfield published an article in *Science* [1] on the use of citation indexes as a new way of classifying journals (at the time he was focussing specifically on the medical literature), one that went beyond the existing keyword- or subject-based classifications systems. He proposed that cited references were more than just acknowledgements of prior research, but were actually links through which one could discover the otherwise hidden conceptual relationships between research areas. He showed how two seemingly disparate areas of scientific or scholarly activity might actually be related because they both cite the same body of prior literature, which in turn might indicate the emergence of a new inter-disciplinary or cross-disciplinary trend.

Dr. Garfield's Science article led him to pursue a grant-funded experimental project to develop a Genetics Citation Index, which was the precursor to the Science Citation Index. Cited reference searching offered a new way for researchers to explore the literature outside their particular fields because they did not need to know the specialized vocabulary of another discipline -- the reference itself was the search term.
Dr. Garfield and other pioneers in the area of bibliometrics and scientometrics (such as MM Kessler and Henry Small) continued their exploration of citation analysis, introducing such concepts as bibliographic coupling, co-citation clustering, citation rankings, and impact factors.
However, Dr. Garfield's vision did not involve merely the intellectual pursuit of an interesting idea, but the practical application of that idea to the everyday world of research and scholarship. His ultimate goal was to provide a new way of exploring the scientific and scholarly literature that researchers themselves could use as a regular part of their research process.

The Science Citation Index marked the first incarnation of that idea, introduced in print in 1964 (followed by the Social Sciences Citation Index® in 1972 and Arts & Humanities Citation Index® in 1978). As in other companies that have been around as long as ISI has, the evolution of

information resources followed the same path: from print, to online, to magnetic tape, to CD-ROM. However, it wasn't until the advent of the Web that the right technology came along in order to fully realise Dr. Garfield's dream.

The Web provided a way to offer a sophisticated method of searching in an easy-to-use interface (in a soon to be ubiquitous environment). With the introduction of the ISI Web of Science® in 1997 (the Web version of the citation indexes) researchers proactively embraced cited reference searching for the first time. Now in 2001 the vision has been expanded to provide a single platform for all the Web-based products offered by ISI.

## The ISI Web of Knowledge

In a nutshell, the ISI Web of Knowledge is a platform encompassing all the content and tools that ISI has developed over the last 40 years, provided in a unified environment for researchers to access directly from their desktops. It includes the evaluated, multidisciplinary content that ISI has always provided -- journal articles, proceedings papers, patents, chemical compounds and reactions, and Web sites and Web documents.

In addition, the platform infrastructure allows us to extend database access by provide complementary sources of information through partnerships with other information companies. The tools provided in the platform include the ones needed to use and manage the content, including subject searching, cited reference searching, browsing, alerting, linking, and exporting.

While the individual sources within the ISI Web of Knowledge continue their separate development path and integration, it is the new technologies that underlie the ISI Web of Knowledge platform infrastructure that are critical to its future development. After all, the Web had been around for a number of years, so what is it that now, in 2001, allowed ISI to actually create a unified research environment? What were the core developments that allowed this platform to take shape? What new technologies are being used to advance database access and linking?

The answer at ISI is two-fold: the introduction of the MuscatDiscovery™ search engine, and the completion of the ISI Links Gateway.

## The MuscatDiscovery Search Engine - its implementation and use in the ISI Web of Knowledge.

Traditionally, ISI has always used Boolean-based search engines, what I think of as "exact match" systems. These systems retrieve exactly what you ask for, which is great if you are like me -- a librarian or researcher who knows how to put together a good search strategy. However, I think everyone knows that this is not always the case with the end-user researcher, and that, frankly, many can use all the help they can get. Because of the platform's modular infrastructure, we can now implement new kinds of technology very quickly and in tandem with an existing technology, and this has allowed us to incorporate the MuscatDiscovery query engine.

Rather than an“exact match”system, the MuscatDiscovery engine is a probabilistic model that uses probability to guess what it is the user wants. I think of this as a“close match”system, where the engine retrieves not only what you ask for but also what it thinks you really meant to ask for. This is the basis of many of the Web search engines currently in use, and is often the tool of choice for the non-search expert. What the MuscatDiscovery engine allows us to do is to expand the ways in which the user can search, so that both structured data (like an A&I database) and unstructured data (like much of the content found on the open Web) can be searched at the same time. Also, the MuscatDiscovery engine can search against documents of different formats, whether they are plain text, HTML, PostScript, or PDF files. Although it retrieves items a bit differently from a Boolean engine -- incorporating the idea of relevance and weighting of terms as it retrieves them -- the nice thing about this engine is that it will still support Boolean-based searching.

There are two ways in which ISI is using MuscatDiscovery engine in the platform: the first is to offer users a mechanism to automatically search for Web documents, and the other is to extend database access by allowing cross-content searching.

## ISI eSearch:

This is the name given to the first implementation of the MuscatDiscovery engine, and it allows users to search the full-text of Web documents indexed from ISI-evaluated Web sites in a way that is completely integrated with journal article retrieval. The way eSearch works is that it takes the Boolean search terms that the user has entered in order to find relevant journal articles, and then translates them“behind the scenes”into a two-part query that is searched against the over 130,000 Web documents currently indexed by ISI editors. The user is then given the option of seeing those documents in a special summary list, with individual items on the list linking directly to the full-text Web document. The entire process extends access to new content in a way that does not disrupt the user's search process by requiring a re-entry of terms or the learning of a new interface.

## ISI CrossSearch

The second implementation of MuscatDiscovery is something we call ISI CrossSearch, which extends searching across different content types -- journals, proceedings papers, and patents -- through a single interface. In addition to offering an easy-to-use“concept”box that requires no knowledge of Boolean command language (researchers can key in a word, phrase, sentence, or cut-and-paste an entire paragraph), it provides a consolidated, de-duplicated results list.

What is particularly important about this cross-content interface is that it allows for a single point of entry into multiple content sources, which means it allows single -sessioning across those multiple resources. In doing so, ISI CrossSearch marks the first phase of the ISI platform.

So the MuscatDiscovery search engine has allowed an expansion and extension of database searching within the ISI platform. Now, what about linking? That brings us to the second technology, the ISI Links Gateway.

## The ISI Links Gateway

Basically, any linking gateway really just needs to answer three questions: Who is the user? Where does the user want to go? And what is the appropriate copy, instance, or version to which we need to send that user? Although the idea seems simple, actual implementation of a robust, comprehensive system that can answer those questions correctly in all different customer scenarios is quite difficult.

In the past, ISI implemented authentication, access, and routing mechanisms the way that many other companies did -- on a product-by-product basis. However, this leads to a level of complexity and duplication that is difficult to manage, and it became clear that what was needed is a single environment for links management across all resources.

This is particularly important in today's collaborative environment in which research libraries are working to build alliances so that resources can be shared across what a number of speakers at this conference have termed the "hybrid library." Our solution was to create ISI Links, a central server which manages all the types of links found across the platform: intra-content links (such as linking from a cited reference to the full article record), inter-content links (such as between journal articles and patents, or between journal articles and gene sequence databases), and links to OPACs. In addition, the gateway allows for context-sensitive linking such as that offered by SFX software, and provides extensive linking to the full-text of articles, whether hosted via Internet by the publisher or hosted locally via Intranet at the customer site level.

Actually, since linking to full-text is such an important issue, I wish to talk more specifically about how we devised our full-text linking system, which we affectionately call "ISI RoboLinks."

## ISI RoboLinks

RoboLinks is the hybrid linking system developed to support a "no dead links" policy.
When we first started working with publishers to establish links to their full-text journal content, we depended on direct electronic feeds sent by the publishers to let us know exactly where a full-text article could be found. Our "no dead links" policy meant that we wanted to make sure that every time a researcher saw a "full-text' button on a record, that researcher could be sure that the full article was available with a simple click. We did not want researchers to hit a dead end (I'm sure we all dread seeing that ominous "404 site not found" message when we click on a promising link), or to be taken to a resource which they are not entitled to use.

Direct publisher feeds ensured stability, but unfortunately it required a great deal of time, effort, and technical savvy to correctly implement the links tables, and although the technical expertise was indeed present, many of the publishers simply were not ready to make the jump along with us.

## Algorithmic Linking

Now there is another type of linking mechanism called algorithmic or "optimistic" links. Optimistic linking involves not a stable link to the full-text of an article, but the storage of meta data

about a full-text article, meta data from which a link can be generated as it is needed by the user. The system stores enough information about an article -- such as the base URL and the algorithm to be used to construct the link -- so that it can attempt to make a link on demand (when the researcher clicks the full text button). This method is relatively easy to employ, but the problem is that you are never quite sure if the electronic article will be found because the link is not actually attempted until the moment you click.

Now ISI RoboLinks is different in that it offers optimistic linking with an added twist -- a measure of dependability. As part of the system the links are actually pre-verified (automatically and on a routine basis) to ensure that the link is still viable. The benefit of RoboLinks is that it allows us to expand our linking system to include the content of publishers who are unable to provide direct feeds, but to do so in a way that ensures a stable system. At present, through a combination of direct electronic feeds and RoboLinks we link to about 2800 journal titles and with agreements to cover over 4,000 titles.

## Conclusion

The first phase of the ISI Web of Knowledge platform was possible because of the way eSearch, CrossSearch, and ISI Links allows individual databases and other resources to be integrated. This leads to the extension of the resources within to those outside the platform as well since the target could be housed outside the confines of this graphical representation. The ways in which the MuscatDiscovery (TM) engine, ISI Links, and other new technologies can be incorporated will allow us to move forward with the next phase of the platform. First, we see an expansion in content sources in specific subject areas, where new partnerships with other information providers allow us to expand database access. We have already partnered with BIOSIS, with CABI Publishing, IEE, and there are others to follow in 2002 and beyond. In addition, the platform's modular infrastructure will allow current tools (such as alerting) to be extended across all content. The ISI Links system will allow expansion of full-text links and the incorporation of context-sensitive linking software such as SFX to other resources (currently only Web of Science is SFX-enabled). I must stress, however, that although there will be ongoing development of the platform, we will not stop enhancing individual resources (for example, set-searching will be available with the next version of Web of Science, allowing a user to combine a cited search with a topic search). The next phase of the ISI platform will be its evolution to a true portal environment, and database access and linking will evolve along with it.

## References

[1] Garfield, E. "Citation indexes for science: A new dimension in documentation through association of ideas." Science, 122(3159): 108-111, July 15 1955.