≪特集：第23回医学情報サービス研究大会≫

インターネット学術情報インデックス（IRI）の構築・運用とネットワーク情報資源

小　山　憲　司*

［抄録］ 東京大学情報基盤センター図書館電子化部門が提供する，インターネット学術情報インデックスの構築・運用を通じて得られた知見をもとに，インターネット上のオープン・アクセス情報について，その特徴と課題について検討した。その結果，オープン・アクセス情報としての学術情報を収集する際には，それを探索する手法と同時に，その内容を判断するための研究者的な視点や主題知識が求められることがわかった。また，それらをデータベースに登録するにあたっては，情報の単位やそのあいまいさへの規則等による対応，バージョン管理などの課題が確認された。また，データメンテナンスの際も，情報の可変性という特徴からさまざまな場面を想定した方法が求められることが明らかとなった。

［キーワード］ インターネット学術情報インデックス，ネットワーク情報資源，オープン・アクセス，学術情報，メタデータ・データベース，サブジェクト・ゲートウェイ，東京大学

はじめに

1999年3月に東京大学でインターネット学術情報インデックス（Index to Resources on Internet，http://resource.lib.u-tokyo.ac.jp/iri/url_search.cgi，以下IRI）（図1）を公開してから，今年で8年目を迎えた。IRIとは，WWWを利用して発信されるオープン・アクセスの学術情報を中心に収集し，これらの情報を効率的かつ効果的に検索・提供できるようにしたメタデータ・データベースである。

IRIの構築は，図書館における情報管理プロセスに類似している。すなわち，世の中にあるさまざまな情報の中から，必要と思われる情報を収集し，それを整理・加工（組織化）し，蓄積・管理（メンテナンス）し，検索・利用に供するというものである。こうしたプロセスの1つひとつを経て，IRIを運営しているのであるが，その情報がいわゆる既存の図書館資料と異なり，インターネット上の情報資源であるため，これら特有のさまざまな問題や課題が発生している。

本稿では，IRIの構築と運営について紹介し，これまで本学が取り組んできた経緯や経験を踏まえながら，ネットワーク情報資源そのものの性質やそれを取り扱ううえでの課題等について検討する。

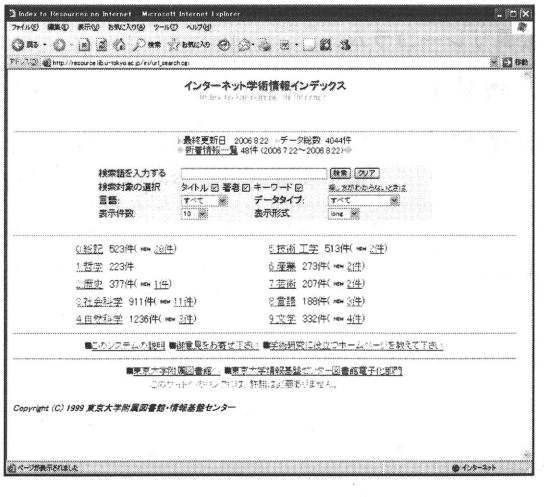

図1　IRIのトップページ

* Kenji KOYAMA
〒101-8430 東京都千代田区一ツ橋2-1-2
国立情報学研究所開発・事業部コンテンツ課

1. IRIの概要

1.1. IRIと情報基盤センター図書館電子化部門

IRIを紹介するにあたり，本サービスを担当するデジタル・ライブラリ係および情報基盤センター図書館電子化部門について触れておきたい。

情報基盤センターは，本学の情報基盤整備を目的として1999年4月に発足した，研究・業務部局である。センターには情報メディア教育部門，キャンパスネットワーキング部門，スーパーコンピューティング部門，そして図書館電子化部門の4つの部門があり，それぞれに教員の所属する研究部門と実務を担当する業務部門がある。

図書館電子化部門は，図書館に軸足を置きつつ，本学の学生および教職員の学習，教育，研究活動にとって必要不可欠な学術情報を提供するための基盤の整備と各種サービスの提供を目的として活動している。業務部門には，主に図書館業務システムの運営・管理を担当する図書館情報係，デジタル化されたさまざまなサービスの利用に係る教育支援を行う学術情報リテラシー係，そしてIRIの担当であるデジタル・ライブラリ係の3係がある。

デジタル・ライブラリ係は，広い意味でオープン・アクセス情報を収集，構築・整理，発信する係といえる。すなわち，①学内で生産された学術情報の収集・整理・発信，②学内（主に総合図書館）で所蔵する資料の電子化，③学内外のネットワーク情報源の収集・整理，および提供といったものが担当する業務としてあげられる。

①は，本学の機関リポジトリ（UT Repository）の構築・運営や学位論文データベースの構築・運用がこれにあたる。②に関連するサービスとして，2006年4月に公開した「鷗外文庫書き入れ本画像データベース」をはじめ，電子版貴重書コレクションの構築・公開などがある。そして，③が今回の報告の主要なテーマとなるサービスであるが，IRIのほか，学内のWebサイトで公開されている学術情報を定期的に収集し，それを自動分類システムによって分類・公開するAcademic Navi U-Tokyoの構築・運用がある。このほか，デジタル・コンテンツを利用するための環境整備として，電子ジャーナルリンク集の整備なども行っている。

1.2. IRIの概要

1.2.1. IRI作成の経緯

インターネットの急速な普及に伴い，Webサイトをはじめ，数多くの情報がインターネット上に公開されているが，その中から，学習や研究に有用な学術情報を検索することは困難な場合が少なくない。たとえば，検索エンジンによる検索ではノイズが混ざったり，仮に検索できていたとしても，表示結果が上位でなかったため，適合情報を見逃してしまうなどの限界があることはよく知られている。また，分野ごと，あるいはWebサイトごとに数多くのリンク集が作成されているが，分野の偏りやその量，検索の方法において限界があるし，何よりも適切なリンク集を探すことが必要になる。

以上のような理由から，学術情報を対象とした，より適切で，有意義なネットワーク情報資源を提供する手段として，IRIを構築し，サービスを開始した。

1.2.2. IRIの現状

2006年8月4日現在，IRIには4,029件の情報が登録されている。その主題別内訳は，図2のとおりである。これらの情報は，キーワード検索とカテゴリー検索の2つの方式によって検索することができる。カテゴリー検索では，日本の標準的な分類表である日本十進分類法（以下，

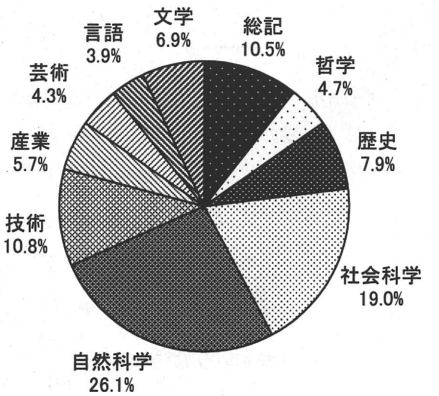

図2　IRIの主題別内訳（2006年8月4日現在）

NDC）を採用しており，分類をたどることで求める情報を検索することができる。

1.2.3. 収録対象

収録対象となる学術情報は，特定の主題に関する書誌・出版情報，各分野に関する総合的な情報リソース，特定主題に関するリンク集，辞書・用語集，データベースなどであり，主として大学・研究機関，学会，学術出版社等が提供する，信頼性の高いサイトを中心としている。

2. IRIにおける情報管理プロセス

2.1. IRIにおける情報管理プロセスの概要

図3は，IRIを構築・運営するにあたって，どのように学術情報資源を収集，組織化し，メンテナンスしているのかについて表した概念図である。簡単にその流れについて，説明する。

IRIに登録したいネットワーク情報資源を見つけたら（図3の①），そのメタデータをIRIのマスターデータベースに登録する（図3の②）。マスターデータベースは，ファイルメーカーProを利用して，パソコン上で管理している。IRIに収録されたメタデータの修正が必要な場合も，まずはこのマスターデータベースを修正・更新する。ネットワーク情報資源の収集状況，更新状況を考慮しながら，マスターデータベースからIRIサーバ用にデータを抽出・変換し，それをIRIサーバに搭載する（図3の③）。利用者は，一般に公開されたIRIを検索し（図3の④），リンクを辿って利用したいWebサイトにアクセスする。

2.2. 情報源の収集と選定

IRIの収集対象となるネットワーク情報資源は，担当職員による探索のほか，学内外からの推薦によってその情報を得ている。一般の利用者と同様，有用なネットワーク情報資源を探索することは，困難な業務の1つであるが，情報を探索するにあたり，次のような情報源を参考にしている。

- 各専門分野の単行書・学術雑誌
- 大学・研究機関，学会，学術出版社等のWebサイト
- 学術情報源を中心としたリンク集
- 上記3点から得られるリンク先

このほか，新聞記事，広報誌・紙など，さまざまな情報源を利用している。

また，収録にあたっては，下記に列挙した「収録リソース選定の目安」を取り決め，これに基づいて選定作業を行っている。なお，収録対象としたWebサイトに管理者への通知を求めるなどの情報が提示されている場合には，基本的にそれを行ったうえで登録している。

1. 内容が学術的で，研究・教育に有用である。
2. 信頼性の高い学術情報源である（典拠など

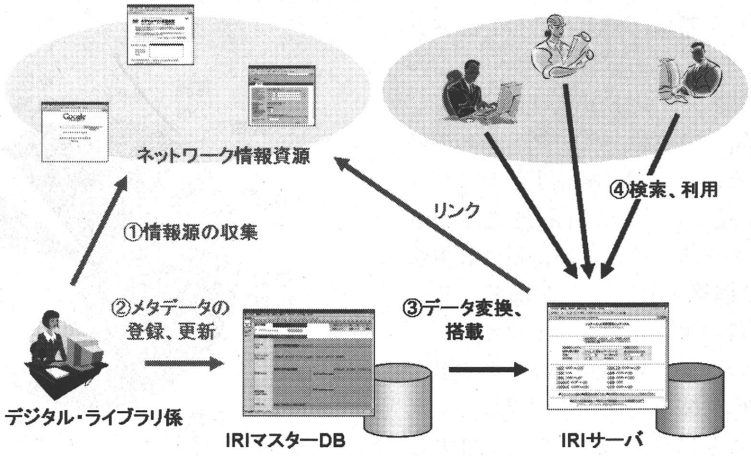

図3　IRIにおける情報管理プロセス

が明示されている等)。
3．継続的に，安定した運営が行われている。
4．量的にある程度充実している。
5．複数の学術情報機関からリンクされている。
6．大学・研究機関，学会，学術出版社等が運営している。
7．企業体のサイトの場合，営利性が低く，内容が研究・教育を中心としている。
8．個人のページの場合，更新頻度が高く，内容が学術的である。
9．個人および特定の団体を喧伝するサイトでない。
10．作者あるいは情報ソースが明示されている(匿名サイトはなるべく採用しない)。

2.3. メタデータの記述と登録

登録するメタデータは，ダブリン・コアを参考にして，書誌記述項目を設定している。その後，国立情報学研究所（以下，NII）のメタデータ・データベースとの互換性を考慮して，項目の追加を行った。このうち，特徴的な項目としては，次のようなものがある。

・ヨミ（タイトル，作成者，責任団体）

日本語によるWebサイトの場合には，ヨミをNACSIS-CATの入力規則に基づいて入力している。

・NDCを利用した分類の付与

分類からの検索を可能とするために，NDCを利用して分類を付与している。いわゆる書架分類と異なり，必要に応じて複数の分類を付与し，多面的な主題検索を可能としている。

・米国議会図書館件名標目表（以下，LCSH）を利用した検索キーワードの付与

統制語による主題検索を実現するために，LCSHを利用している。

・キーワードの付与

LCSHによる統制語のほか，日本語によるキーワードも付与している。

・データタイプNIIの付与

NIIが提供しているメタデータ・データベースで必須項目となっているデータタイプに応じて，付与している。

3. IRIとネットワーク情報資源

ここからは，これまでのIRIの構築・運用の経験から，オープン・アクセス情報としてのネットワーク情報資源の特徴やそれを取り扱う際の課題，問題点について，収集，組織化，そしてメンテナンスの3つの側面から検討する。

3.1. 収集における課題

ネットワーク情報資源の収集にあたっての一番の課題は，何度も触れているが，広範なインターネット上の情報の中から，いかにそれらを探索し，収集するかということである。GoogleやYahoo!をはじめとする検索エンジンのように，強力な検索手段が用意されている現在，いかに有用な学術情報を適切かつ豊富に提供できるかは，われわれにとって大きな課題の1つである。そのためには，手がかりとなる情報源を複数おさえるとともに，ロボット検索などの新たな手法を開発する必要があるのかもしれない。

また，収録対象候補となる情報が見つかったとしても，次に求められるのが，それをIRIに収録すべきか否かの判断である。現在は，先に掲げた「収録リソース選定の目安」によって収録の可否を判断していることがほとんどであるが，学術上あるいは研究上有用な情報資源を収集・選択するためには，研究者としての視点や，各主題分野の知識も必要となるだろう。

3.2. 組織化における課題

収録対象となったネットワーク情報資源をIRIに登録するにあたって問題となるのは，大きく次の3つである。1つは，何を組織化の対象とするか。2つ目は，何を記述の情報源とするか。そして最後に，どのように記述するか，である。

3.2.1. 何を組織化の対象とするか

これは，収録の対象となる情報の単位（ユニット）が明確ではないことに起因する。たとえば，土木学会が公開している「土木デジタルアーカイブス」(http://www.jsce.or.jp/library/page/report.html)は，「戦前土木絵葉書ライブラリー」のような画像データベースをはじめ，複数のデータベースの目次の役割を果たすWebページである。この場合，個々のデータベースを収録対

象とするだけでなく，その目次にあたる当該ページも収録することが望ましいと思われる。

例にあげたWebサイトをはじめ，一般にネットワーク情報資源は，複数のファイルとそれらを結ぶリンクによって構成される。このときに，ファイルの集合を1つの情報資源とみなすこともできるし，個々のファイル，あるいはそのなかの一部が収録にふさわしいと判断する場合もあるだろう。そのような場合，組織化の対象をどの範囲に設定するかは，ネットワーク情報資源を扱う際に考慮すべき特徴の1つといえる。

3.2.2. 何を記述の情報源とするか

IRIでは，ダブリン・コアを参考にして，書誌記述項目を設定しているが，それを記述するためにどこを参照するのかは規定していない。したがって，個々のネットワーク情報資源を確認しながら，メタデータを記述することになる。しかし，ネットワーク情報資源の記述言語の代表であるHTMLは，マークアップ言語であるものの，そこに記述される情報やその使い方は作成者側に大きく委ねられている。そのため，たとえば，titleタグによって示された情報をそのままその情報資源のタイトルとするかどうかの判断に迷うことも少なくない。このほか，作者・作成機関が明示されていないなどの責任性（著者性）のあいまいさや，公開日などの出版事項のあいまいさといったことも，ネットワーク情報資源を記述する上で留意すべき特徴の1つであろう。

さらに，ネットワーク情報資源の場合，情報の可変性という特徴を考慮する必要がある。Webサイトのなかには，更新が行われず，情報が固定化されて公開されているものもあるが，一般にその内容は常に変化している。こうしたヴァージョン管理は，現在のIRIの運用体制ではほとんど不可能である。また，IRIでは，Identifierに格納されたURLを，個々のWebサイトを同定識別するためのエレメントとして利用しているが，URLが同一であり続けるからといって，その内容が同じであるという保障はどこにもない。URLが同じでも内容がまったく変わってしまうこともあり得るということも注意すべき事項の1つである。

3.2.3. どのように記述するか

この課題は，3.2.1.や3.2.2.で取り上げた事例からもわかるように，記述規則，すなわち記述項目および記述の方法を明確にすることが解決策になる。しかしながら，IRIでは，1999年のサービス開始以来，3.2.1.や3.2.2.のような課題を抱えながらも，書誌記述項目やデータ管理の方法について，見直しや更新をほとんど行ってこなかった。特に，データ管理については，XMLへの対応なども未着手のため，他システムとのメタデータの相互利用やRSSによる情報発信など，新たなサービスへの対応ができていないのが現状である。このことは，IRI独自の問題であるが，ここに記しておく。

3.3. メンテナンスにおける課題

3.2.2.でも触れたが，ネットワーク情報資源の特徴の1つに，情報の可変性がある。そのため，収録している情報について，定期的なメンテナンスが必要となる。IRIでは，フリーウェアのリンクチェックソフトを使って，リンク切れやURLの変更などについては，メンテナンスを行っているが，収録した情報そのもの（メタデータ）についての確認や修正は行っていないのが現状である。

リンクチェックは，毎月1回定期的に行っている。具体的には，まずリンクチェックソフトにより，収録された情報のURLに接続できるかどうかをチェックする。エラーが出たものについては，別途手動で確認を行う。エラーが出たものの中には，直接URLを指定すれば接続できるものもあれば，日をおいて確認すれば接続できるものもある。このほか，既存のURLだけでは確認できないものは，検索エンジンなどを使ってURLの変更の有無を確認したりする。

さて，図4は，2006年5月に実際にリンクチェックを行った結果である。対象となったのは，3,938件で，このうちエラーとなったものは321件で，全体に占める割合は8.2%であった。逆に，問題のなかったものは3,617件で，91.8%にあたる。エラーのあったもののうち，問題のなかったものは171件で，残り150件（3.8%）が何らかの問題があった。その内訳は，URL変更が

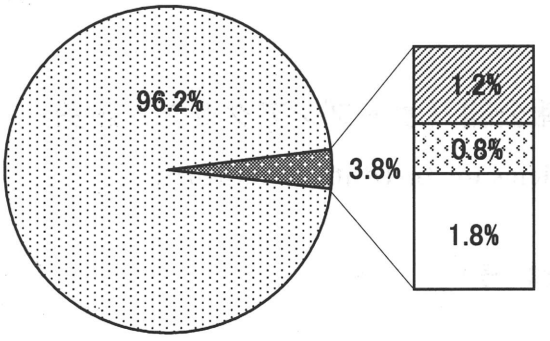

図4　リンクチェック結果

49件（1.2%），サイト消滅などが31件（0.8%），一時的にアクセスできないと思われるサイトが70件（1.8%）であった。

なお，3.2.2.でも指摘したように，リンクチェックソフトによってエラーにならなくても，その内容が変わっている場合も十分考えられるが，それは今のところ確認することができない。また，Webサイトによっては，URLの変更を通知する際に，元のURLのWebページにその旨記載しているものも見受けられる。この場合，接続できるかどうかをソフトウェアを使って自動的にチェックする現在の方法では，URLの変更を確認することができない。リンク切れを確認するなど，単なるリンクチェックだけでは不十分であることも留意しなければならない。

おわりに

本稿は，本学のIRIでの取り組みを通じて得られた知見などから，オープン・アクセス情報であるネットワーク情報資源の特徴やその取り扱いについての課題，問題点を検討することが目的であった。インターネットが急速に普及し，そこで発信・公開される情報がこれだけ身近になった現在においては，言わずもがなのことも多々あったかと思われる。また，取り上げた内容やその方法が不適切であったため，検討が十分でなかったものもあったかとも考えられる。これらの点については，今後の課題としたい。

最後に，ここで得られた検討結果をフィードバックし，よりよいサービスを提供できるよう，業務に生かしていきたい。

なお，本稿は，2006年7月15日から16日に千葉大学で開催された，第23回医学情報サービス研究大会・継続教育コースにおいて，筆者が発表した内容を加筆・修正したものである。研究大会ならびに本誌において，IRIについて紹介する機会をいただき，改めて本学が提供するサービスについて，検討することができた。関係者の皆様にここに記して謝意を表したい。また，本稿を執筆するにあたり，東京大学情報基盤センターデジタル・ライブラリ係の赤津さんに資料の準備等を手伝ってもらった。この場を借りて，お礼申し上げる。

参考文献

1) 米田寿宏. 東京大学における電子図書館サービス：インターネット学術情報インデックスを中心に. 情報の科学と技術. 50 (10), 2000, 510.
2) 大川直子ほか. 東京大学附属図書館におけるインターネット学術情報インデックスの作成について. 大学図書館研究. 56, 1999, 12-22.
3) 栃谷泰文. ゲートウェイ・サービスのためのメタデータ：「インターネット学術情報インデックス」作成の事例報告. 現代の図書館. 38 (1), 2000, 55-62.
4) 栃谷泰文. “ゲートウェイサービスのためのメタデータ：「インターネット学術情報インデックス」作成の事例報告”. 電子資料の組織化：日本目録規則（NCR）1987年版改訂版第9章改訂とメタデータ. 東京, 日本図書館協会, 2000, p. 57-71. (ISBN 4-8204-0003-7)
5) 東京大学情報基盤センター図書館電子化部門. (インターネット), 入手先<http://www.dl.itc.u-tokyo.ac.jp/>, (参照 2006-08-21).

（原稿受付け：2006.8.21）