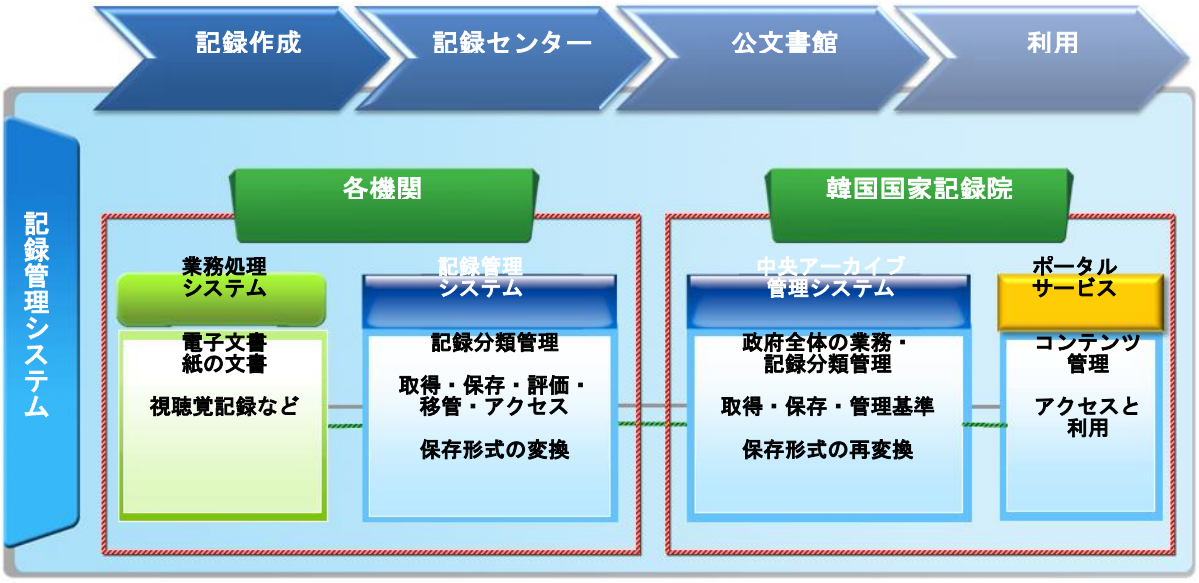
韓国のウェブアーカイビングの現状と見通し

ソ・キョンナン
韓国国家記録院

1. 電子記録の形式

大韓民国公文書管理法（Public Records Management Act of the Republic of Korea）に従い、電子記録は電子文書、ウェブアーカイブ（ウェブサイト）、および行政情報データセットの3種類に分けられる。

電子文書については、2000年代の初頭に電子文書作成システムが構築された。このシステムに基づき、各部局で作成された記録は一定期間を経て記録センターに移管される。その後、永久保存に値する記録は、一定の期間記録センターで保存された後、永久保存機関である韓国国家記録院（以下、NAK）に移管される。現在、2000年代初頭に作成された電子文書をNAKに移管しているところである。



＜韓国の電子記録管理システム＞

行政情報データセットについては、試行プロジェクトの下で暫定的なシステムが構築されているが、多様な形式のデータセットを標準化した上でNAKに移管することは容易では

ない。そのため、行政情報データセットを移管するための研究は現在進行形で継続中である。

2. ウェブアーカイビング

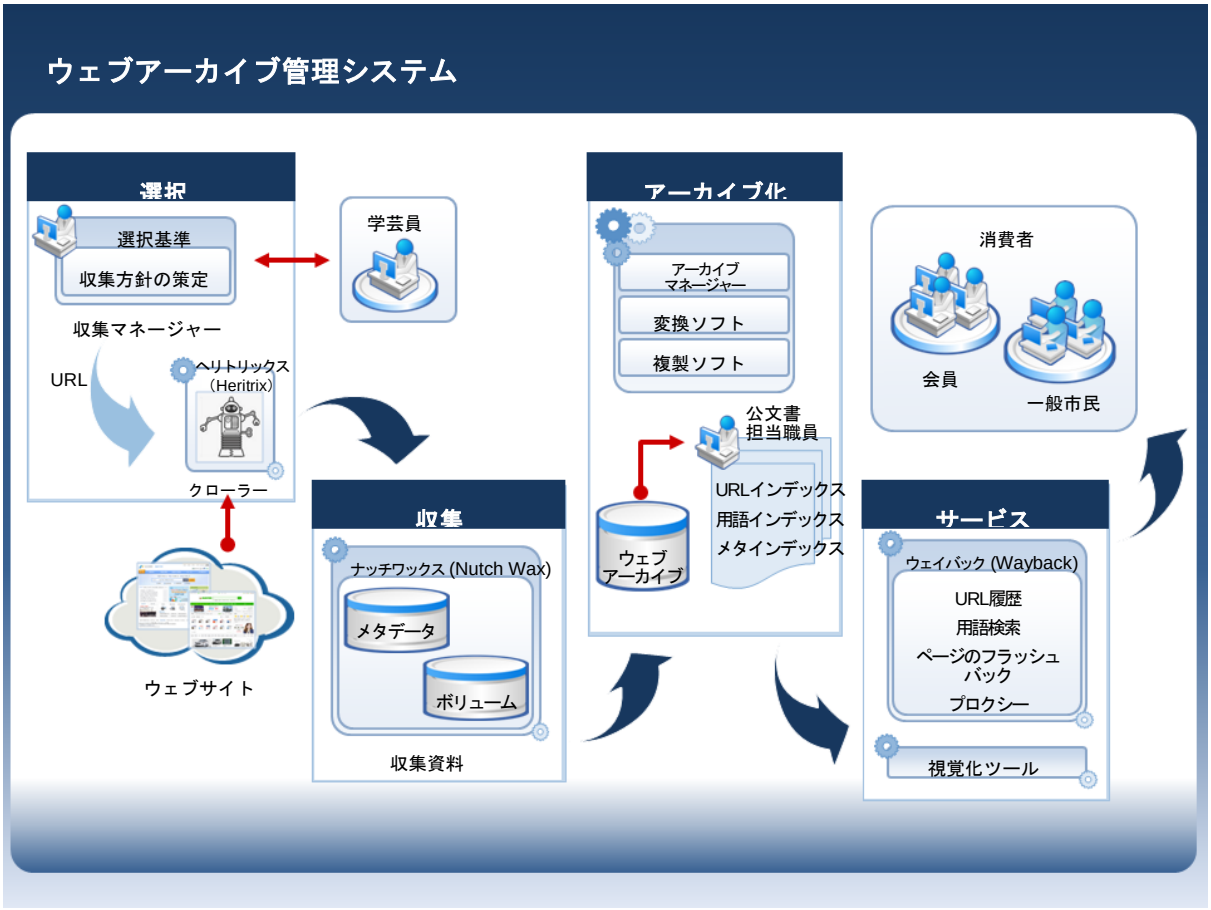
ウェブアーカイビングは、ワールド・ワイド・ウェブからその一部を収集し、集めたものをアーカイブの形で保存し、そのアーカイブへのアクセスを可能にして記録を利用できるようにするまでのプロセスである。過去と比べるとデジタル形式で作成される記録は多く、ホームページ、ブログ、およびSNSなどで利用に供される記録は飛躍的に増えている。ウェブ上の記録は量的に増えている半面、そうした記録の収集に対する意識は十分に喚起されていない。

ウェブ上の記録は変わりやすく、短命で、不安定である。つまり、簡単に更新されてしまい、ライフサイクルが短いために次世代まで残らない。貴重な情報を掲載する主要なウェブサイトが更新され、時には消えてしまい、その結果として情報にアクセスできなくなってしまうことは、非常に残念なことである。

特に、公共機関のウェブサイトは、政府と一般市民の間の意思の疎通を図る役割を担っている。韓国政府は大統領選挙の結果を受けて5年ごとに代わり、この変化に伴って政府部局の再編や統合が行われる。廃止された部局のウェブサイトが適切なタイミングで収集され、保存されなければ、こうしたウェブサイトに掲載された情報は将来利用できなくなってしまう。

3. 韓国のウェブアーカイビングの現状

韓国では2004年以降、韓国国立図書館がOASIS（Online Archiving & Searching Internet Sources）というプロジェクトの下で、一部の中央政府機関、大学、電子刊行物などの主要ウェブサイトの収集を行ってきた。NAKにおいては2008年に調査プロジェクトを実施し、2011年にウェブアーカイブ管理システムを構築した。2008年の調査プロジェクトではアメリカのインターネットアーカイブ、オーストラリアのPANDORA（Preserving and Accessing Networked Documentary Resources of Australia）、日本の国会図書館など海外のさまざまな事例を分析し、その結果として、ウェブ記録コレクターや複製・表示装置などの基盤技術と標準形式を確立した。NAKはウェブアーカイビング用の標準やツールを開発しているIIPC（International Internet Preservation Consortium）が提供するツールやソフトウェアやWARC形式を採用し、それらをNAKのウェブアーカイブ管理システムに応用している。



＜NAKのウェブアーカイブ管理システム＞

NAKはヘリトリックス（Heritrix）というウェブクローラー【訳注：ウェブサイトの自動収集ソフト】を使ってウェブアーカイブを収集し、それをWARC形式でストレージに保存する。保存されたウェブアーカイブは、ウェイバック（Wayback）装置と名付けられた複製・表示装置を使って利用することができる。

このシステムを構築して以来、NAKは約50の公共機関のウェブサイトを収集し、保存してきた。

4. 主要課題

他のデジタル分野と同じように、ウェブ環境は日常的に変化と消滅を繰り返している。ここで、ウェブアーカイブ管理の観点から論議を呼んでいる以下のような問題について説明したい。

4.1 包括的収集 vs. 選択的収集

包括的収集は、選択の手順をまったく踏まずに、特定のURLの下にあるすべてのウェブアーカイブを収集することと定義される。その特徴は迅速なプロセスにあり、また、ウェブサイトの前後のつながりを完全に保存することができる。しかし、ウェブアーカイブのストレージと保存にかかるコストが相対的に高い。

選択的収集は、あらかじめ定めた機関別指針に沿ってウェブアーカイブを収集することであり、集めるべきアーカイブの量と質を自動収集ソフトに詳しく理解させる。多額のコストが発生し、時間もかかる。さらに、重要なウェブアーカイブを取り込むための指針が主観的で、かつ限定される。そのため、収集範囲や収集方針から外れた記録は、将来、消えてしまう危険がある。そうなれば、次世代はそうした記録を利用することができなくなる。

現在のところ、オーストラリアとイギリスと日本が選択的収集の方針を採用しており、アメリカとスウェーデンとフィンランドは包括的収集の方針を採用している。NAKは包括的収集の方針を採用しているが、どちらの方針が保存に適しているかという問題については、依然として意見が分かれるところである。

4.2 直接収集 vs. リモートハーベスティング

直接収集は、対象機関に依頼して、ウェブアーカイブのコピーを物理的にアーカイブの中へ移管してもらう収集方法である。この場合、プログラムソース、データベースに関するデータ、環境設定など、復元するために必要なすべての情報を収集または捕捉（キャプチャ）する必要がある。復元に使うソフトウェアのライセンスが必要になるため、多額のコストがかかる可能性がある。

リモートハーベスティング【訳注：遠隔操作による収集】は、ウェブクローラーまたはウェブロボットのソフトウェアを使って、リモートサーバーからウェブアーカイブを捕捉することである。自動ツールを使うのでコストが相対的に小さくて済む。しかし、質の観点から見て、ウェブアーカイブの捕捉における精度が相対的に低い。

NAKはウェブアーカイブの捕捉には基本的にリモートハーベスティングの方針を採用しており、一部、遠隔操作による収集が不可能な場合に直接収集の方針を採用している。しかしながら、大統領のアーカイブについては、主に直接収集の方針を採用している。現在のところ、NAKは第14代から第18代までの大統領に関連するウェブサイトを収集し、一般向けにサービスを提供してきたが、サービスコストの増加という問題に直面している。そのため、今後は大統領アーカイブについてもウェブクローラーのソフトウェアを使うべき

か検討中である。このように、直接収集とリモートハーベスティングは、機関の状況やネットワーク環境条件を考慮しながら同時に使うことができる。そのため、長所と短所を天びんにかけた上で、調整可能な方針を選ぶことになる。

4.3 収集サイクル

前述の通り、ウェブアーカイブは往々にして変更されやすく、消去されやすいものである。そのため、ウェブアーカイブ捕捉の収集サイクルは、対象機関の種類、ウェブサイトの規模、変更頻度などを総合的に検討した上で決めるべきである。

韓国の場合、5年ごとに政府の交代が繰り返されることを考慮して、中央政府機関のウェブサイトは政権交代の前と後に捕捉され、一方、外郭機関のウェブサイトは5年に1度捕捉される。また、ウェブサイトの規模により、収集にかかる期間は1日から数カ月までばらつきがある。さらに、アジア大会や国際博覧会のように一般市民が関心を寄せる国際的なイベントについては、ウェブサイトの変更を確認するために収集間隔を週1回や月1回に短縮する場合もある。このように、収集サイクルを決定するためには、ウェブサイトのさまざまな特性を考慮すべきである。

4.4 収集できない記録

デジタル時代の主な特徴は共有化（sharing）と開放性（openness）である。他国の政府と同じく韓国政府もまた、政府の透明性と信頼性を高めるためあらゆる取り組みを行っている。韓国政府は、公共情報の積極的な共有化と省庁間に存在する障壁の撤廃を促進し、協力関係を改善していくための政府運営の新たな枠組みとして、「政府3.0」（Government 3.0）と名付けた政策を進めている。しかし、こうした取り組みにもかかわらず、ウェブ開放性を指数で評価した調査結果によると、韓国の指数はまだまだ高くない。評価項目の中には、「ウェブクローラーによるサイトへのアクセスがブロックされているか否か」というものがある。独自の方針に従い、ウェブクローラーがウェブサイトを捕捉することを認めない機関があると、ウェブアーカイビングは実施できない。そこでNAKは、「ウェブクローラーによるアクセスを拒否」という設定を解除するように対象機関に勧告することで、問題なくウェブアーカイビングを進めている。

この事例とは別に、たとえ最新版のウェブクローラーを用いたとしても、技術的に捕捉や収集ができないウェブサイトがある。たとえば、収集にあたりログインする必要があると、ウェブクローラーにはログイン操作が不可能なため、そのウェブサイトにアクセスできなくなる。さらに、現在では、ダイナミック・ウェブ・アプリケーションを実現するた

めにAjax（Asynchronous JavaScript＋XML）が使われることが多いものの、Ajaxのような技術はW3Cの定める標準にはなっていないので、そうしたウェブサイトをクローラーで捕捉することは難しい。つまり、ウェブブラウザーの設計方式や実行方式が異なると、ウェブサイトの収集や表示が困難になるのである。また、URLを動的に生成するJavascriptを使用しているウェブサイトや、Flashに別のリンクがある場合は、JavascriptまたはFlashにリンクされたURLをウェブクローラーが抽出できないので、ウェブアーカイブの収集が不可能になる。こうした事例のほか、問題なく収集が完了しても、復元や表示の段階で何らかの問題が生じる場合も多い。ウェブ標準に対応した技術を使うよう対象機関に勧告しているにも関わらず、ここで説明したいくつかの事例はヘリトリックス（Heritrix）というウェブクローラーの側の技術的制約に起因するものであり、したがって、将来、さらなる技術的発展が果たされるべきである。

5. 今後の見通し

韓国の電子政府は、国連の電子政府評価において3期連続で第1位にランク付けされたことでよく知られている。しかし、多様な形式に対処するための電子記録管理方式が、量的および質的な情報の拡大に追いつかない。特に、他の技術と比べて軽視されているウェブアーカイビング技術を改善するためには、専門職員や専門家の育成と、関連する調査や研究の推進を続けていくべきである。

NAKは、前述のようなウェブアーカイブ管理に関連する問題を徹底的に検討することを通じて、ウェブアーカイブ管理政策を補強していくつもりである。加えて、収集範囲と収集サイクルを広げるためには、追加的なシステム拡張が必要である。というのも、現在のところ、ウェブアーカイビングの対象が中央政府機関の一部に限られているからである。さらに、NAKは捕捉、収集すべきウェブアーカイブの保管・保存のために、ウェブアーカイブ・コレクターおよび安全なストレージの導入に必要な予算を手当てすべきである。

NAKは次世代のためにウェブアーカイブを保存するためだけでなく、それらを調査や教育の資料として利用するための施策も策定すべきである。また、NAKは不朽の価値を持つ民間の貴重なウェブアーカイブを収集することにも関心を寄せており、新たなサービスモデルを開発するため、あらゆる取り組みを行うべきである。