組織内 Web アーカイブシステムにおけるアーカイブデータの整合性について

柊和佑，阪口哲男，杉本重雄
筑波大学図書館情報メディア研究科
〒305-8550 茨城県つくば市春日 1-2
E-mail: {ragi, saka, sugimoto}@slis.tsukuba.ac.jp

概要

近年は様々な組織が Web ページのアーカイブに取り組んでいる。先に我々は、Web ページの発信者と Web アーカイブが連携することで、収集と提供を効率良く行う組織内 Web アーカイブを提案した。通常、一度アーカイブされたデータはそのまま変更されずに保存し続けることが求められるが、組織内 Web アーカイブでは何らかの変更の必要性が生じる場合がある。しかし、アーカイブデータは制限無く変更をしても良いものではない。そのため本稿では、変更を行う際に保つべきアーカイブデータの整合性について考察し、この整合性に反しない Web アーカイブの変更について考える。

キーワード: 組織内 Web アーカイブ，アーカイブ内容の変更，アーカイブデータの整合性，イントラネット

1 はじめに

現在、膨大な数の情報資源が Web ページとして発信されている。その内容は日々作成・更新され、それに伴って消えてしまうものも多い。そのため、近年は様々な組織が Web ページのアーカイブに取り組んでいる。

我々が提案した組織内 Web アーカイブのモデルとそのシステム(Institutional Web Archive System：IWAS）は、運用者と発信者が収集と提供の条件を指定し、組織内で運用する Web アーカイブである[1]。条件となるメタデータは、Web ページの運用組織が必要に応じて定義した独自のメタデータである。そのため、IWAS で構築されたアーカイブデータは、運用組織毎にメタデータが異なり、運営組織に改組が起きた場合、そのまま利用することが困難だった。そこで、我々は改組の際に、複数の Web アーカイブが協調的に動作し、他の Web アーカイブのデータを受継ぐためのモデルとその機能を定義した[2]。

その課程において、アーカイブデータを受け継ぐために、一部のアーカイブデータの削除が必要となることが判った。このような変更は Web アーカイブの運用の中でも行われる可能性がある。しかし、Web アーカイブのデータを制限無く変更することは Web アーカイブの役割を損なう可能性がある。そこで本稿では、IWAS におけるアーカイブデータの整合性を検討する。また、変更については一部のアーカイブデータを削除する場合を考える。

2 章では我々がすでに提案している組織内 Web アーカイブと、論文[2]で提案した IWAS の統合手法を述べる。3 章では、IWAS の整合性の定義を行い、統合後にその定義で整合性が取れているかを検討する。そして、変更後の整合性の維持について、削除を例に考察する。4 章では今後の展望について述べる。

2　組織を指向した Web アーカイブ

2.1　IWAS の概要とその構造

我々の提案した組織内 Web アーカイブ[1]では、Web ページを作成している発信者と、IWAS を運用している運用者が連携して、各組織の責任の範囲内で Web ページの収集・蓄積・提供を行う。組織内で Web アーカイブを構築するため、網羅性の高い Web アーカイブを構築することができる。IWAS の主な機能は収集、蓄積、利用があり、発信者および運用者が定めた条件に従って動作する。この条件を記述したものをポリシー記述と呼び、このポリシー記述には収集ポリシー記述(Collection Policy: CP)と利用ポリシー記述(Access Policy: AP)がある。

収集時、管理機能は CP に従って Web リソースの特定と収集の可否を判断し、収集を実行する。その後、CP に基づいて XML で記述したメタデータの付与を行い、収集日時の異なる同一 URL のリソースにまとめて蓄積する。利用時は、蓄積されたデータのメタデータと閲覧者のリクエストを AP に照らし合わせ、利用の可否を判定する。 [Fig2.1]

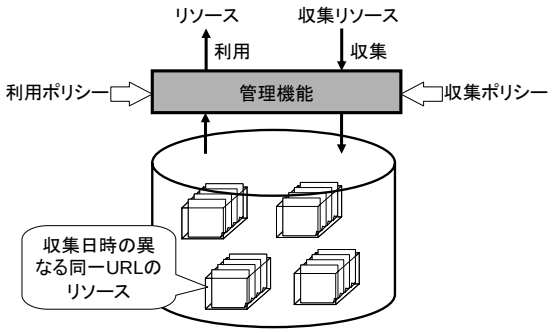

Fig2.1　IWAS 概要

2.2　WAO の構築と Collection Policy

IWAS では蓄積したデータを Web Archive Object(WAO)と定義し、WAO は以下の4つの要素から構成される。

- Component：URL と日時で識別される Web ページを構成する最小単位。発信者が定めたメタデータが付与される。付与されたメタデータは、その Component で構成される Resource のメタデータとして参照される。また、Component は1つ以上の Resource に属する。
- Resource：収集した時点の Web ページを表すオブジェクト。収集した時点の Web ページを構成する Component がメタデータに記述されている。また、Resource のメタデータには、発信者が指

定した Component(Main Component)のメタデータを利用する。また Resource には1つ以上の Main Component が存在する。

- Resource Set：同一 URL で示される異なる Resource をまとめたもの。その URL について、使用開始日時と使用停止日時が記録されている。
- Access Policy：利用ポリシー記述。

そして、WAO は Fig2.2 のように表すことができる。

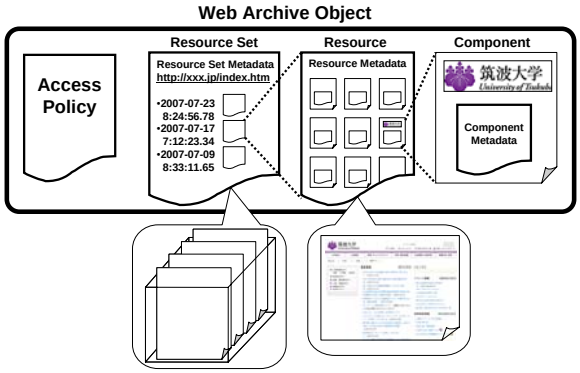

Fig2.2　Web Archive Object Model

IWAS に蓄積された Component には、IWAS を運用する組織が必要に応じて定めた記述規則に従ってメタデータが記述されている。このメタデータは、Web ページの収集条件とメタデータ付与のルールを発信者が CP として記述し、それに基づいてシステムが付与する。このとき、収集条件である URL を、正規表現を使ってまとめて記述することで、発信者の記述する量を減らしている。これは、それぞれの Web ページが内容によってまとめて管理されており、そのまとまり毎に URL を一定のパターンで示すことができるからである。

さらに、一つの WAO 内のオブジェクトは全て同じ CP を使って作られるという性質がある。そのため、運用者や発信者が CP を変更する毎に、変更後の CP で構築された WAO が Fig2.3 のように新たに構築される。そのため、CP によって WAO が何をアーカイブしたのかが判るのである。

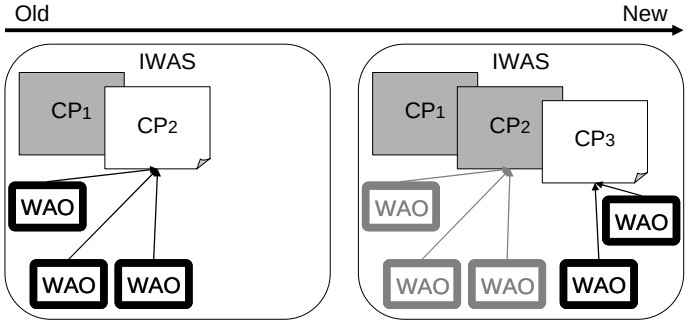

Fig2.3　Collection Policy と WAO の関係

2.3　IWAS における改組への対応

IWAS において改組に対応するため、論文[2]の中では、我々は組織の改組の種類をあげ、その解決法として他の Web アーカイブのデータを受継ぐためのモデルとその機能を定義し、統合と名付

けた。Fig2.4 のように統合を行うためには、メタデータスキーマの変換と、統合する WAO の選別が必要である。統合の際には、大量の WAO を一度に変換する必要があるため、そのための変換規則を統合ポリシー記述(Merge Policy: MP)として統合システムに与えることで行う。

MP は WAO の中から必要なデータだけを選別する。これは、統合後の IWAS にとってはそのデータは削除されたことと同じである。しかし、MP にはその変換規則があるため、統合後の WAO は MP を見ることによって、統合時にどのデータがなぜ削除されたかがわかるものとなっている。そのため、前述した CP と合わせれば、WAO とアーカイブ対象であった Web ページとの関係が明確になる。

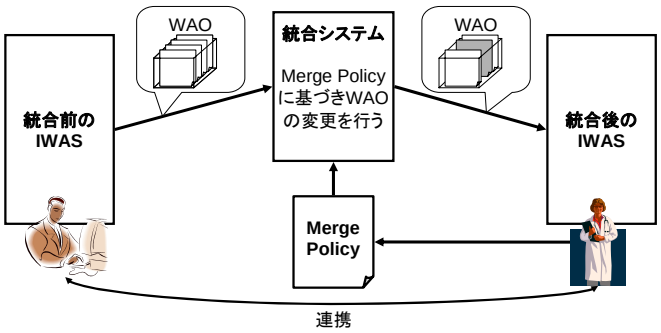

Fig2.4　統合システム概要

2.4　ディジタルアーカイブのデータの変更

ディジタルアーカイブのデータを変更することに関しては、いくつかの論文で議論されている。ミラーは、論文[3]において、プライバシー保護のための原則として情報の死について述べている。この論文ではプライバシーを確保するためには削除され死をあたえなければならない情報がある、と論じている。また、収集対象は Web ページではないが、各国で行われているディジタルアーカイブや機関リポジトリにおいては、いったん公開された情報の実体や、そのメタデータの変更をどのように認めるか議論が行われている[4]。

実際に発信者の要請に従ってアーカイブの変更を行っているものとしては、Internet Archive[5]がある。Internet Archive では、要請に従い、公開の停止か、アーカイブからの削除を行っているほか、発信者が robot.txt を用いて収集を拒否した場合は、過去に収集した分も遡って削除している。

IWAS でも、何らかの制約の元、役割を損なわない範囲の変更は必須の機能であると思われる。

3　組織内 Web アーカイブにおける整合性

3.1　Web アーカイブの整合性

IWAS が役割を果たすためには、IWAS のアーカイブ内容が整合性のとれた状態でなければならない。これは、IWAS のアーカイブ対象である Web リソースが、その収集開始時から現在まで矛盾無く蓄積されている状態である。そのためには、WAO が何をアーカイブしたか判るようになっていて、利用者がその情報を参照できなければならない。我々は、その情報が CP と MP であると考える。

前述したように、IWAS では、アーカイブデータは CP に基づいて構築される。そして、構築する際

の状態によって、IWAS のアーカイブ対象となった Web リソースは以下のように分けられている。

(1) 収集され、アーカイブされた
(2) 収集時には存在したが、何らかの理由で収集されなかった

この場合のうち、(1)は収集時の Web ページを正確にアーカイブしている場合であり、利用することができるアーカイブデータである。(2)は、運用者や発信者のポリシーによって Web アーカイブに蓄積されなかった Web リソースで、CP によって収集時に対象外であったことを利用者に示すことができる。現在の IWAS では、IWAS に存在しない理由を記録している。

統合以前の IWAS では、個々の Web リソースが (1)または(2)のいずれに該当するかが CP によって判別される。そして、統合が起きた場合、Web リソースは(3)のように状態になる。

(3) 収集されたが、何らかの理由でアーカイブデータが削除された

(3)の状態になっているアーカイブデータは、CP によって収集対象であることが、MP によって統合時に削除されたことがわかるようになっている。

つまり、収集時に用いた CP と、統合時に用いた MP を残すことで、IWAS がアーカイブデータの状態を利用者に説明することができる。このように、収集時の Web リソースと、この状態を表す記述と、WAO の状態が一致していることを、本稿では IWAS の整合性が保たれているとする。

3.2 変更と Web アーカイブの整合性

統合以外の変更を WAO に加える場合でも、実際に変更された WAO の他に、CP や MP のような記述を別途残すことで整合性を維持できると考えられる。本節では、個別の WAO に変更を加える際、どのような情報を残せば IWAS の整合性を維持できるか、削除を例に考察する。

まず、削除を行う場合、そのリクエストは削除対象を URL と、日時範囲で指定すると考えられる。これは、リクエストを出すのがその Resource を閲覧した利用者であり、利用者はアーカイブデータを Web ページという単位で閲覧、認識しているためである。そして、IWAS は時間的な幅をもつデータを扱っているため、そのリクエストには時間的な幅が指定される。この幅があるため、削除のリクエストは複数の WAO にまたがる可能性がある。

また、Resource を指定して削除を行う場合、その Resource を構成する Component 全てに影響が出る。この Component 群は、一部が他の Resource を構成している可能性があり、単純に消すわけにはいかない。さらに、IWAS では Resource のメタデータを、Main Component のメタデータを参照することで取得しているため、Resource のもつ一部の Component 群の情報を削除することはできない。

以上から、IWAS におけるデータの削除には以下の情報が残っていれば CP、MP、と共に利用することで、整合性が保たれると考えられる。

- 削除リクエストの対象を示す URL と日時範囲

- Resource とそれを構成する Component の情報
- Main Component のメタデータ

まず、リクエストには URL と日時範囲があるため、どの WAO のどの Resource を削除するかはわかる。しかし、リクエストは各 Component の URL を示してはいないため、リクエストだけではどの Component を削除したのかは判らない。そこで、Component を、Main Component のメタデータだけを残して削除し、Resource の持つ各 Component の情報を残す。これにより、どの Component がどの Resource に対するリクエストで削除されたのかが判る。ただし、他の Resource を構成している Component の場合は、削除対象 Resource のもつ Component の情報を削除し、Component を残す。

以上により、削除された Resource に対して利用のリクエストがあった場合はその Resource が削除されたことを示すことができる。そして、削除された Component に対して利用のリクエストがあった場合でも、どの Resource に対するリクエストによって削除されたかがわかる。これは、IWAS のアーカイブ内容にどのような変更がなされたか記録してあり、さらに利用者がそれを参照することができる状態であるといえる。

以上のように、IWAS では変更のリクエストおよび、変更されたアーカイブデータの情報を記録することで、アーカイブ内容の整合性を保つことができる。

4 まとめと今後

本稿では、Web アーカイブを実運用する場合にアーカイブの変更がおきても、整合性を維持するための方向性を示した。Web アーカイブにおいて、アーカイブ内容の変更は見落とされがちではあるが、必須の機能である。実際に運用する際は、発信者から変更のリクエストが出される可能性が高く、その際にどう対応するか決めておく必要がある。本稿では削除の場合について考えたが、今後は削除以外の変更作業を含めて考察をさらに進め、IWAS を Web アーカイブとしての役割を保ちつつ、柔軟に運用できる組織内 Web アーカイブのモデルとしたいと考えている。

参考文献

[1] Wasuke Hiiragi, Tetsuo Sakaguchi, Shigeo Sugimoto, Koichi Tabata: “A Policy-based System for Institutional Web Archiving”, Z.Chen et al.(Eds):ICADL 2004,LNCS 3334, pp.144-154,2004.

[2] 柊 和佑, 阪口哲男, 杉本 重雄: “分割・統合可能な組織内 Web アーカイブシステムの構成方法”, 情報知識学会論文誌, 2008 年掲載予定

[3] Miller, Arthur R: “コンピューターとプライバシー”, 鶴田尚美訳. 情報倫理学研究資料集 第 1 集. 水谷雅彦 編. 京都, 京都大学文学研究科「情報倫理の構築」プロジェクト室, 1999, p. 105.

[4]高木和子: “世界に広がる機関レポジトリ: 現状と諸問題”, 情報管理 2004 , 47(12), p. 806-817.

[5] Internet Archive. <http://www.archive.org/>. (accessed 2008-2-25)