1L-02

アーカイブ横断型メタデータ連携による東日本大震災アーカイブ群からのコンテンツ集約手法

積 佑典† 本間 維‡ 三原 鉄也‡ 永森 光晴††‡‡ 杉本 重雄††

筑波大学情報学群情報メディア創成学類† 筑波大学大学院図書館情報メディア研究科‡

筑波大学図書館情報メディア系†† 知的コミュニティ基盤研究センター‡‡

1. はじめに

東日本大震災発生後、多数の震災アーカイブが作成・公開された。資料が非常に多く蓄積公開され、個別資料を検索できる反面、検索結果には類似する資料が大量に並んでしまうといった課題がある。また、アーカイブ毎にメタデータの記述内容が異なるため、横断的な検索に課題がある。本研究では、こうした点に注目し、日付や場所、内容といった複数のメタデータ記述項目を利用し、資料を集約することによってアーカイブ資料へのアクセス性を高める取り組みを進めている。本稿では、これまでのメタデータの分析から得た震災アーカイブのメタデータの特徴と実験的な資料集約の結果を報告する。

2. 震災アーカイブの問題点

複数の震災アーカイブから資料を横断的に検索できるポータルサービスとして「NDL 東日本大震災アーカイブひなぎく」[1]がある。このサービスでは現時点で 300 万件を超える大量の資料から検索を行うことができるが、必要な資料を的確に絞り込むことが難しく、検索結果には類似する資料が多数に並び、多様な資料を参照することが難しい。また資料の主題や意味内容まで十分に記述することができておらず、資料の作成意図を把握するのが困難であるといった問題がある。

本研究はこれらの問題を解決するため、類似・関連する資料を集約して検索・閲覧する方法の開発を目的としている。資料の集約により、ユーザに見せる類似した資料数が減り、一度に多様な資料を見せることができる。また、個々の資料では作成意図や価値がわかりづらかったものがより明確化する。

"A Method of Metadata Aggregation from Digital Archives of the Great East Japan Earthquake"
†Yusuke Seki. School of Infomatics. Univ of Tsukuba.
‡Tsunagu Honma. Tetsuya Mihara. Graduate School of Library, Information and Media Studies. Univ of Tsukuba
††Mitsuharu Nagamori. Shigeo Sugimoto. Faculty of Library, Information and Media Science. Univ of Tsukuba.
‡‡Research Center for Knowledge Communities. Univ of Tsukuba.

3. 震災アーカイブからのメタデータ収集

ひなぎくと連携している震災アーカイブのメタデータは、いずれも NDLKN 形式[2]で記述されている。記述内容には、資料の作成背景・主題などに関する情報が含まれている。本研究では、OAI-PMH を利用して、青森震災アーカイブ[3]、久慈・野田・普代震災アーカイブ[4]、みちのく震録伝[5]、河北新報震災アーカイブ[6]、全 4 アーカイブのメタデータを取得した。

4. メタデータの特徴分析

本研究では、4 アーカイブから収集したメタデータを対象としてメタデータ作成の時間推移をカテゴリ・共起キーワードの観点から調査した。

4.1. カテゴリ別時間推移

メタデータとして「カテゴリ」の情報が与えられている青森震災アーカイブ、久慈・野田・普代震災アーカイブに関して、カテゴリ別の資料作成数の推移を調査した。両アーカイブ共に「復旧・復興」「被害」など全 13 カテゴリに分けられている。

震災発生からの時間経過によって資料作成数は減少するものの、作成数の累積で見た場合、各カテゴリの割合は「復旧・復興」が増えていき、「被害」は減っていく（図 1）。これは直感的にはごく自然な遷移である。

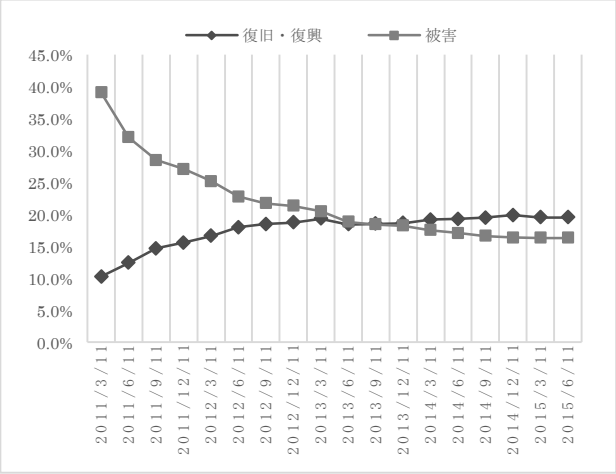

図 1 カテゴリ別資料作成数累積の割合
（久慈・野田・普代震災アーカイブ）

表1　4アーカイブで出現する共起キーワードとその出現頻度（上位5件）

語1	語2	青森	久慈	震録	河北	合計
東日本大震災	被害	2	9	1	9,597	9,609
津波	被害	3	177	116	8,637	8,933
建物	自動車	1	1	5,423	11	5,436
地震	被害	3	61	12	5,298	5,374
地震	津波	35	307	1	5,019	5,362

4.2. 資料間共起キーワードとそのアーカイブ間共起キーワード別時間推移

個々の資料にはキーワードが付与されているが、個別のキーワードでは資料の意味を詳しく表現することができない。メタデータ内で同時に出現する二つのキーワード（共起キーワード）を利用することで、より詳しく資料を表現できると考え、全アーカイブで出現する共起キーワードを持つメタデータ数を調査した。この調査により、アーカイブ毎にキーワードの付与に特徴があることがわかった（表1）。

5. 実験的な集約作成

本研究では集約作成の実験段階として、まず資料の作成場所・作成日・作成者のメタデータでの集約を行った。次に、より小さい集約を作成するためのメタデータの調査を行い、タイトル・シリーズタイトルを含めて集約を行った。

5.1. 作成場所・作成日・作成者による集約

ここでは、単純に同値の作成場所・作成日・作成者をメタデータとして持つ資料群を一つの集約とした。この集約作成後、集約ごとにメタデータの調査を行った。集約に含まれる資料の数（集約サイズ）毎の集約数の割合は、アーカイブ毎に差が出た（表3）。また、この3項目を用いた集約では集約サイズが1,000を超えるものもあり、さらに小さな集約を作成する必要があることがわかった。

5.2. タイトル・シリーズタイトルによる集約

メタデータを分析したところ、青森震災アーカイブ、久慈・野田・普代震災アーカイブにはシリーズタイトルが付与されており、また、一連の資料には同一タイトルが付与されるというアーカイブ個別の特性がわかった（表2）。これを基にタイトル・シリーズタイトルで集約を行い、サイズの大きい集約の数を減らすことができた。その反面、集約サイズ1の集約も多数作成され、これらの集約のうち関連するものを一つに集約しなければならない。

また、キーワードによる集約も検討したが、アーカイブによって付与傾向が異なったため、アーカイブ毎に異なる手法でメタデータを扱わなければならず、将来の課題としている。

表2　集約に使用するメタデータをもつ資料数

メタデータ(総資料数)	青森(68,297)	久慈(127,500)	震録(124,132)	河北(115,861)
GPS	68,297	126,379	95,938	0
地名	66,820	0	96130	98,975
作成日	61,879	117,071	124,025	114,636
作成者	68,296	127,500	124,076	112,635
タイトル*	11,637	20,853	850	84,134
シリーズタイトル*	66,750	127,497	0	0

*ユニークなタイトル数・シリーズタイトル数

表3　集約サイズ別集約数（作成場所・作成日・作成者）

アーカイブ	100以上	100未満2以上	1
青森	202(1.4%)	5,059(35.8%)	8,860(63.7%)
久慈	215(1.2%))	5,491(30.4%)	12,345(68.4%)
震録	507(1.3%)	35,076(87.5%))	4,489(11.2%)
河北	91(0.1%)	23,524(35.8%)	39,238(62.4%)

表4　集約サイズ別集約数（作成場所・作成日・作成者・タイトル・シリーズタイトル）

アーカイブ	100以上	100未満2以上	1
青森	122(0.4%)	6,768(20.9%)	25,570(78.8%)
久慈	159(0.4%)	7,560(17.1%)	36,481(82.5%)
震録	412(0.8%)	42,362(79.5%)	10,517(19.7%)
河北	26(0.0%)	12,380(8.1%)	141,226(91.9%)

6. おわりに

今回の集約実験では、多数の資料群から小さいサイズの集約を作成した。現在、作成した集約の精度の評価を進めている。小さい集約を合わせて大きい集約を作成するために、メタデータのさらなる特徴分析・加工を検討する必要がある。アーカイブ毎に異なる特徴を持つという調査結果に基づき、アーカイブ毎にヒューリスティックなメタデータ加工が必要である。また、作成された集約が新たなコンテンツとして利用されるべく、集約に対するメタデータの付与も必要である。

将来的には、資料や作成した集約はLinked Data技術を用いて震災アーカイブ内外との横断的なメタデータ連携を行い、Linked Open Dataとして公開することが望まれる。

参考文献

[1]NDL東日本大震災アーカイブひなぎく. http://kn.ndl.go.jp/ (参照 2015/12)
[2]NDL東日本大震災アーカイブメタデータスキーマ(2014年9月版). http://kn.ndl.go.jp/sites/default/dfsfiles/ndlkn_schema_jpn.pdf. (参照 2015/12)
[3]青森震災アーカイブ. http://archive.city.hachinohe.aomori.jp/ (参照 2015/12)
[4]久慈・野田・普代震災アーカイブ. http://knf-archive.city.kuji.iwate.jp/ (参照 2015/12)
[5]みちのく震録伝. http://search.shinrokuden.irides.tohoku.ac.jp/shinrokuden/ (参照 2015/12)
[6]河北新報震災アーカイブ. http://kahoku-archive.shinrokuden.irides.tohoku.ac.jp/kahokuweb/search/ (参照 2015/12)