CD-005

Web Index による Web ページ中の構造化データの二次利用の提案
Proposal on Reuse of Structured Data in Web Pages by Web Index

大島 拓也† 遠山 元道†
Takuya Ohshima Motomichi Toyama

1. はじめに

著者らは，所属する研究室のプロジェクトとして Web Index システム（以下，WIX システム）の研究および開発に取り組んでいる．WIX システムは，Web ページの閲覧者の主導による情報資源の統合を実現するシステムである．

WIX システムでは，Web ページの閲覧者が，その目的に応じて Web ページ中の単語をハイパーリンクに変換し，その単語に関連する情報のある Web ページに素早くアクセスすることができる．図 1 に，WIX システムによるハイパーリンクへの変換例を示す．Web ページ中の単語について，赤色になっている部分にハイパーリンクが生成され，「田中将大」をクリックすれば田中将大に関する情報が，「ヤンキース」をクリックすればヤンキースに関する情報にアクセスすることができる．WIX システムはクライアント-サーバシステムとして構築され，クライアントは Safari や Google Chrome, Mozilla Firefox などの PC 用 Web ブラウザの拡張機能[1]や，携帯端末向け iOS にバンドルされた Safari の拡張機能[7]として実装されている．

ハイパーリンクの変換は，WIX ファイルに基づいて行われる．WIX ファイルは，XML 形式のファイルであり，WIX エントリが複数記述されている．WIX エントリはキーワードとターゲット URL の組として記述されている．キーワードは変換の対象となる単語を示しており，ターゲット URL はどのような Web ページへのハイパーリンクに変換するかを示している．図 1 の例では，キーワードに田中将大，ターゲット URL に田中将大に関する Wikipedia の記事の URL(https://ja.wikipedia.org/wiki/田中将大)を指定した WIX ファイルを利用しており，この場合には，田中将大というアンカーテキストを持つハイパーリンクをクリックすることで Wikipedia の田中将大の記事にアクセスすることができる．

WIX システムの中心的なコンテンツである WIX ファイルの作成方法については，これまで様々な方法が提案されてきている．その多くは，Web ページなどの情報資源から抽出したい情報を選択し，WIX ファイルを生成するものであった．これは特定の Web サイトを対象にする場合には向いているが，複数の Web サイトから同じような情報を集約する場合でも，Web サイトごとに抽出する情報の選択が必要となり，WIX ファイル作成者の負担が大きくなるものであった．

本論文では，WIX ファイルの作成について，Web ページに記述された構造化データを収集し，そのデータから情報を抽出，集約することで複数の Web サイトにまたがる大規模な WIX ファイルを作成するシステムについて提案し，その実装について述べる．本研究は，提案するシステムによって生成された WIX ファイルが，WIX システムの利用価値を高めることを目的とする．

†慶應義塾大学大学院理工学研究科 Graduate School of Science and Technology, Keio University

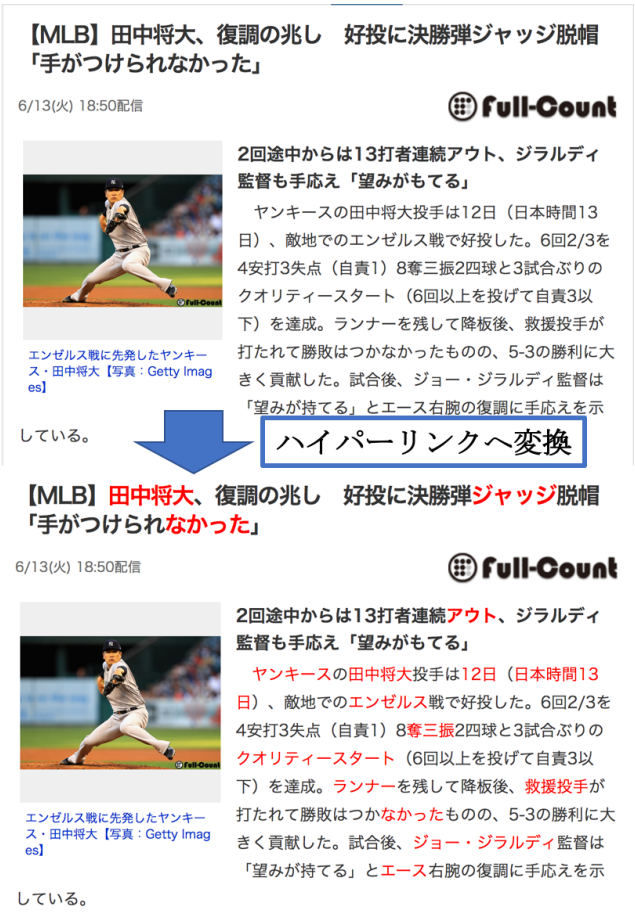

図 1 WIX システムによるハイパーリンクへの変換

本論文の構成は以下の通りである．まず 2 章で本研究の背景となる WIX システム，WIX ファイルの作成方法と構造化データについて説明する．3 章では本研究のために設計した WIX ファイル生成システムの実装について述べる．4 章で評価について言及し，5 章でまとめを行う．

2. 研究の背景

2.1 WIX システム

2.1.1 WIX ファイル

WIX ファイルは，XML 形式で記述されたキーワードとターゲット URL の組み合わせである WIX エントリの集合である．エントリには，キーワードとなる見出し語を keyword 要素として，それに対応する Web ページの URL を target 要素として記述する．header 要素にはファイルの概要，作者のコメント，WIX ファイルの言語を記述する．body 要素にエントリを記述する．WIX ファイルの記述例

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE WIX SYSTEM "http://wixdemo.db.ics.keio.ac.jp/wixfile.dtd">
<WIX>
  <header>
  …
  </header>
  <body>
    <entry>
      <keyword>すし匠</keyword>
      <target>https://s.tabelog.com/tokyo/A1309/A130902/13000852/</target>
    </entry>
    <entry>
      <keyword>うどん 丸香</keyword>
      <target>https://s.tabelog.com/tokyo/A1310/A131003/13000629/</target>
    </entry>
    <entry>
      <keyword>パティスリー・パリ セヴェイユ</keyword>
      <target>https://s.tabelog.com/tokyo/A1317/A131703/13005198/</target>
    </entry>
  </body>
</WIX>
```

図 2 WIX ファイルの記述例

を図 3 に示す．WIX ファイルは，「Wikipedia の見出し語と対応する記事の集合」や「レストラン名とそれに対応するレストラン紹介サイトの記事の集合」など，意味を持つエントリの集合として作成する．

2.1.2 ハイパーリンクの生成処理

WIX システムのサーバは，WIX サーバとして構築される．WIX ファイルの作成者は WIX サーバに WIX ファイルを登録する．登録されたファイルは関係データベースに展開され，その情報を元に Find Index と呼ばれるオートマトンを構築する．オートマトンは Aho-Corasick 法に基づき辞書式マッチング処理を行うものであり，その処理時間はオートマトンのサイズに関わらず入力された Web ページの文字列長にのみ比例する[2]．

WIX システムのクライアントは，ブラウザ拡張機能として実装されている．ユーザがクライアントをインストールすると，ツールバーが表示される．ツールバーに登録しておいた WIX ファイルを指定すると，閲覧中の Web ページの HTML が WIX サーバに送られ，オートマトンによるマッチングが行われる．WIX サーバはマッチングした WIX エントリの情報をレスポンスとして返し，その情報を元にクライアントは HTML を書き換え，ハイパーリンクが生成される．

2.2 WIX ファイルの作成

WIX ファイルの作成について，もっとも単純な方法は人間の手によって直接記述する方法である．この方法は，「ある大学の特定の学科に所属する研究室の名前とその Web サイト」に基づき，数十件のエントリを含む WIX ファイルを作成する場合など，データ数が少ない場合には利用可能であるが，データ数が大きくなると作業量が大きくなる．

より大規模な WIX ファイルを作成する場合，Web 上の情報資源からデータを抽出，加工して作成する方法が提案されてきた．Web ページからハイパーリンク集合を発見して自動生成する方法[3, 4, 5]，RSS ファイルから自動生成する方法[6]などがある．

また，一般的な Web スクレイピングの方法でデータを抽出し，加工することで WIX ファイルを作成することも可能である．代表的な公開 Web サービスとして import.io[16]がある．また，Ducky[8]のようにブラウザライクな GUI で抽出する属性を指定することができるものもある．

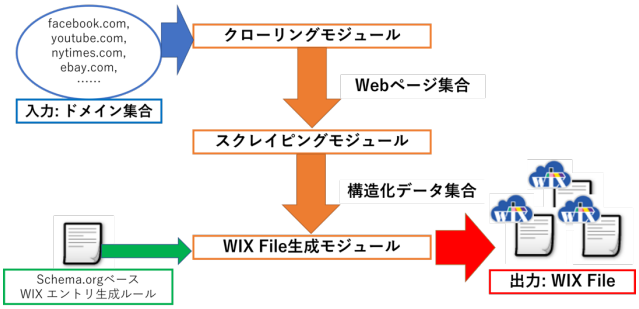

図 3 自動 WIX ファイル生成システムのアーキテクチャ

これらの方法を利用した WIX ファイルの作成は，いずれの場合もデータの抽出元となる Web ページや Web サイトの構造に対応して，データの抽出方法を指定する必要がある．つまり，WIX ファイルの作成者は，1 つのページに 1 つの抽出方法，1 つのサイトに 1 つのクローリング方法を与える必要がある．これは，複数のサイトからデータを抽出して大規模な WIX ファイルを作成する際に大きな負担となる．

2.3 構造化データ

2013 年頃から，Web ページへの構造化データの記述が増加している[9, 10]．構造化データは，Web ページに記述された情報について，検索エンジンやその他のクローラーが利用できるように記述したものである．

構造化データの利用例として，Google Rich Snippets[17]がある．これは，Google の提供する検索エンジンの検索結果に，レストランであれば評価や価格帯，イベントであれば日程や会場などが一緒に表示されるというものである．

現在，幅広く利用されている構造化データのシンタックスとしては，Microdata[11]，JSON-LD[12]や RDFa[13]の三種類がある．これらはいずれも W3C によって推奨されている．

また，構造化データのセマンティクスを与える取り組みとして，Schema.org[14]がある．Schema.org では，Restaurant や Book など，Web ページに記述される様々なエンティティとそのスキーマが与えられている．スキーマとして，Restaurant であれば address や priceRange，Book であれば author や publisher などの記述可能なフィールドが定義されている．

これらのシンタックスとセマンティクスの組み合わせによって構造化データが記述されている．構造化データは Web サイト間で共通となるため，例えば，異なるレストラン紹介サイトから同じデータ構造を持つ Restaurant の構造化データを抽出することが可能になる．

3. 自動 WIX ファイル生成システム

3.1 システムの提案

本研究では，構造化データを利用した自動 WIX ファイル生成システム（以下，本システム）の提案を行う．

本システムでは Web ページ中に記述された構造化データを大規模に収集し，収集したデータから WIX エントリの生成ルールに基づいて情報を抽出，集約して WIX ファイルを生成する．本システムの利用によって，WIX ファイルの作成者はサイトごとに抽出する属性の指定を行う必要

がなくなり，少ない労力で大規模な Web サイトを対象にした WIX ファイルを作成することが可能となる．

3.2 システムの外部仕様

本システムのアーキテクチャを図 2 に示す．

入力となるファイルには URL を 1 行あたり 1 つ記入する．URL としてドメイン名を記述した場合は，robots.txt を参照し，利用可能であれば sitemap.xml に基づいたクローリングを行う．ドメイン名以外の URL を記述した場合，ドメイン名を記述したが sitemap.xml が利用可能でない場合は，Web ページ中に含まれるハイパーリンクをたどってクローリングを行う．

WIX エントリの生成に関しては，スキーマごとに WIX エントリの生成ルールを定義する．スキーマ Organization とスキーマ NewsArticle で定義した生成ルールの例を＆＆に示す．ルールには，Schema.org のどのスキーマを利用するかを指定する type フィールドと，WIX エントリのキーワードおよびターゲットを指定する keyword フィールド，target フィールドがある．keyword フィールドや target フィールドには，各スキーマで定められているフィールドの名前を指定する．また，target フィールドについては，構造化データそのものには直接含まれていない情報である，構造化データの存在している Web ページの URL を，”#url”として指定することができる．WIX ファイルの target には，通常この”#url”を指定する．1 つのルールを設定すると，全ての Web ページから取り出した構造化データを同じルールに基づいて WIX エントリに変換する．したがって，同一のスキーマを持つ構造化データについて，複数の Web サイトのデータが混在した WIX ファイルを作成することが可能となる．

3.3 システムの内部仕様

入力された Web サイトのドメインに基づき，まずクローリングモジュールで Web ページのクローリングを行う．クローリングを行う際には，利用可能であればクローラーに対するアクセスのルールを規定する robots.txt を参照し，そこから sitemap.xml を取得して効率的なクローリングを実現する．クローリングの結果として，Web ページの集合を得る．

スクレイピングモジュールでは，Web ページの集合から構造化データを抽出する．本モジュールでは，2.3 節で紹介した 3 つのシンタックスのうち，特に幅広く利用されている Microdata と JSON-LD の 2 つについて実装を行なった．

JSON-LD の抽出では，”<script type=”application/ld+json”>…</script>”として HTML 中に記述された部分を，JSON としてそのまま抽出する．

表 1 Microdata の要素の種類と対応する値

要素の種類	値
meta	content 属性の値
audio, embed, iframe, img, source, track, video	src 属性から展開した絶対 URL
a, area, link	href 属性から展開した絶対 URL
object	data 属性から展開した絶対 URL
data, meter	value 属性の値
その他	その要素の値

```
organization = {
        “type” : “Organization”,
        “keyword” : “name”,
        “target” : “#url”,
}
news_article= {
        “type” : “NewsArticle”,
        “keyword” : “headline”,
        “target” : “#url”,
}
```

図 4 WIX エントリ生成ルールの記述例

Microdata の抽出では，HTML 要素を示すタグ中に宣言された itemscope 属性を発見し，同じタグに併記されている itemtype 属性からスキーマを取得する．itemscope の宣言された要素の内側にある，itemprop の値をフィールド名とし，その値は itemprop 属性の含まれるタグに基づき，表 1 の通りに決定される．ただし，itemprop 属性を含む要素が itemscope 属性を含んでいる場合は，その itemscope 属性が定義するアイテムを値とする．これにより，構造化データでアイテムを階層的に定義することができる．これらは W3C によって勧告された Microdata の仕様に基づくものである．

WIX ファイル生成モジュールでは，抽出した構造化データから，特定のスキーマを持つものについてデータを抽出し，WIX エントリの生成を行う．

抽出された WIX エントリは，スキーマごとにまとめられて 1 つの WIX ファイルとなり，これが最終的な出力となる．10 スキーマに対して WIX エントリ生成ルールを実装すれば，10 個の WIX ファイルが生成される．

4. 評価

4 章では，まず構造化データに基づくデータの抽出の有効性を調べるために行った構造化データの実装調査について示し，続いて本システムで生成された WIX ファイルについての評価を行った．

4.1 構造化データの実装調査

2.3 節で述べた通り，構造化データの実装については 2013 年頃からその記述が増加していることが明らかになっている．一方で，本システムで利用可能な形式の構造化データがどの程度実装されているかについては明らかになっていない．そのため，本システムで利用可能な構造化データの実装状況について調査を行なった．

調査は，SEO に関連する情報を提供している moz.com[15]の”The Moz Top 500”に基づき行なった．この Web ページでは，対象ドメインへのリンク数の多いドメインをランク付けし，上位 500 ドメインについて示したものである．参照されることの多いドメインで構造化データが実装されていれば，本システムがそこから WIX ファイルの生成に必要な情報を抽出することが可能となる．

調査は，2017 年 6 月 1 日の”The Moz Top 500”からドメインのランキングを取り出し，各ドメインについて，自作のスクレイピングツールを用いて，トップページからハイパーリンク 2 つ以内で参照可能な Web ページについて，本システムで利用可能な構造化データが含まれているかを判定することで行なった．本システムで利用可能な構造化データとは，シンタックスに Microdata か JSON-LD のいずれかを利用し，セマンティクスに Schema.org を使用している構造化データである．

表 2 スキーマごとの構造化データ実装ドメイン数

スキーマ名	ドメイン数
Organization	25
ImageObject	13
Person	12
NewsArticle	11
WebPage	6
WebSite	6
BreadcrumbList	5
Offer	4
SearchResultsPage	4
VideoObject	4
AggregateRating	3
ListItem	3
Product	3
Rating	3
CollectionPage	2
InStock	2
PostalAddress	2
AudioObject	1
Brand	1
ItemList	1
MusicRecording	1
MusicAlbum	1
MusicGroup	1
SpeakableSpecification	1
WPHeader	1
SiteNavigationElement	1
WebPageElement	1
WPFooter	1
SoftwareSourceCode	1
LocalBusiness	1
Restaurant	1
EducationEvent	1
Event	1
Place	1
SoftwareApplication	1

調査結果を図 5 に示す．本システムで利用可能な構造化データの記述されたドメインは 273 個あり，上位 500 ドメインの半数以上で構造化データが利用可能だと確認できた．構造化データの利用可能なドメインの例として，facebook.com, youtube.com, nytimes.com, ebay.com, adobe.com などがある．一方で，170 ドメインでは構造化データが利用できなかった．構造化データの利用できないドメインの例として，amazon.com, instagram.com, baidu.com, fc2.com などがある．また，アクセスすると他のドメインにリダイレクトされるドメインが 46 個存在した．このグループに属するドメインとしては，goo.gl, youtu.be, t.co などがある．最後に，アクセスを拒否されるドメインが 11 個存在した．このグループに属するドメインは，miitbeian.gov.cn や wixsite.com があり，これらは www.miitbeian.gov.cn や www.wixsite.com など，特定のドメインを指定しないとアクセスすることができなかった．

4.2 生成された WIX ファイルの評価

”The Moz Top 500”の上位 100 ドメインのうち，構造化データが利用可能な 48 ドメインについて，本システムを用いて WIX ファイルを生成し，評価を行なった．

スキーマごとの構造化データ実装ドメイン数を表 2 に示す．ただし，灰色の網掛けの行については，構造化データの存在は確認したものの，BreadcrumbList や

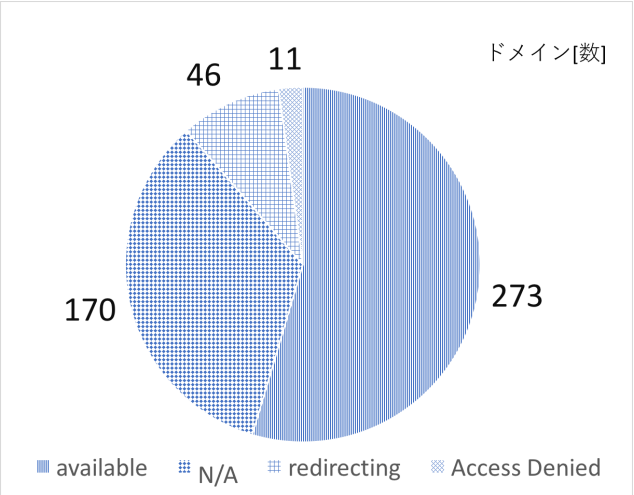

図 5 構造化データの調査結果

SearchResultsPage など，Web ページの構造や部品を定義するスキーマや，ImageObject や AggregateRating など，インスタンスが名前を持たず，WIX ファイルによるハイパーリンクの生成に不適切なスキーマを示している．これらについては WIX エントリ生成ルールの実装を行なっておらず，WIX ファイルも生成されない．表 2 によると，Organization スキーマは調査対象としたドメインの半数以上で実装されていることがわかる．図 4 に示したルールに基づき，Organization，NewsArticle スキーマの情報を抽出して作成した WIX ファイルの一部分を表 3 に示す．

本システムの実装には問題点もある．例えば，表 3 に示した Organization の WIX ファイルでは，Apple というキーワードに対し，apple.com ドメインの Web ページがターゲット URL である WIX エントリが複数登録されている．抜粋していない本来の WIX ファイルでは，このようなエントリが数万件も登録されている．これは，Apple の公式サイトで，全ての Web ページに Apple に関する構造化データが登録されていることによって生じた問題である．このような問題は Apple 以外でも，特に企業や組織の公式サイトを対象にした場合に生じる．

表 2 に示したスキーマごとの構造化データ実装状況は，全世界で被リンク数の多い上位 100 ドメインについて調査したものである．このため，被リンク数の少ないドメインが持つ構造化データの実装はこの表には現れない．そのようなドメインとスキーマの例として，食べログ(tabelog.com)やぐるなび(gnavi.co.jp)の Restaurant, 技術評論社 (gihyo.jp) や翔泳社 (shoeisha.co.jp) の Book，楽天(rakuten.co.jp)や Wowma!(wowma.jp)の Product などがあげられる．これらの Web サイトからも，構造化データを抽出して WIX ファイルを作成することは可能である．

5. おわりに

本研究では，Web ページへの記述が近年増加している構造化データを利用した WIX ファイルの作成方法について提案し，Schema.org をセマンティクスに，JSON-LD と Microdata をシンタックスに取っている構造化データからの WIX ファイル生成システムについて実装および評価を行なった．本システムの利用によって，Web サイトごとの要素

指定を排除し，大規模な WIX ファイルの自動生成が実現する．

参考文献

[1] 林昌弘，青山峻，朱成敏，遠山元道. “Keio WIX システム (1) ユー ザインターフェース”. データ工学ワークショップ，DEIM2011. 2011.
[2] 森良介，藪達也，朱成敏，遠山元道. “Keio WIX システム (2) サーバーサイド実装”. データ工学ワークショップ，DEIM2011. 2011.
[3] 市東隼，分部亮太，朱成敏，遠山元道．”Keio WIX システム (3) コンテンツ作成”．DEIM2011．2011．
[4] 藤井洋太郎，遠山元道．”WIX システムにおけるコンテンツ作成 支援システム”．DEIM 2012．2012．
[5] 金岡慧，遠山元道．”自動更新型 WIX ファイル生成システムおよ び Deep Web に対するアタッチ機構の構築”．DEIM2014．2014．
[6] 大島拓也，遠山元道．”Web Index における配信型コンテンツを利用した自動ライブラリ更新システムの提案”. DEIM 2016. 2016.
[7] 生田史織，遠山元道．” iOS プラットフォームにおける Web Index システムの提案”．DEIM 2016, 2016.
[8] Kei Kanaoka, Motomichi Toyama. “Browser GUI for generating web data extraction rules in Ducky”. iiWAS No.79, pages 1-5, 2015.
[9] Robert Meusel, Christian Bizer, Heiko Paulheim. “A Web-scale Study of the Adoption and Evolution of the schema.org Vocabulary over Time”. WIMS No.15, pages 1-11. 2015.
[10] Robert Meusel, Petar Petrovski, Christian Bizer: “The WebDataCommons Microdata, RDFa and Microformat Dataset Series”. Semantic Web Conference vol.1 pages 277-292. 2014.
[11] “HTML Microdata”. W3C. https://www.w3.org/TR/microdata/
[12] “JSON-LD 1.0”. W3C. https://www.w3.org/TR/json-ld/
[13] “RDFa Core 1.1 – Third Edition”. W3C. https://www.w3.org/TR/rdfa-syntax/
[14] “Schema.org”. Schema.org. http://schema.org/
[15] “The Moz Top 500”. Moz, Inc. https://moz.com/top500
[16] “import.io”. import.io. https://www.import.io
[17] “Search Features”. Google Developers. https://developers.google.com/search/docs/guides/search-features

表 3 生成された WIX ファイルの例

Organization の WIX ファイル

Keyword (name 属性)	Target (構造化データの存在するページの URL)
Lady Gaga	https://www.facebook.com/ladygaga/
TOYOTA	https://www.facebook.com/ToyotaMotorCorporation/
Apple	https://www.facebook.com/apple/
Apple	https://www.apple.com
Apple	https://www.apple.com/privacy/privacy-policy/
Apple	https://www.apple.com/watch/
Weebly	https://www.weebly.com/
Jimdo GmbH	https://www.jimdo.com
AOL	https://www.aol.com/article/news/2017/06/27/house-speaker-paul-ryan-defends-the-senate-republican-health-car/23005193/
AOL	https://www.aol.com/article/news/2017/06/27/teen-accused-of-running-over-killing-best-friend-while-high/23004969/

NewsArticle の WIX ファイル

Keyword(headline 属性)	Target (構造化データの存在するページの URL)
Discovery of crash victim’s body hours after wreck shocks cops	https://www.yahoo.com/news/m/a6904ec7-db81-315b-a205-b32850d321c6/ss_discovery-of-crash-victim%E2%80%99s.html
Repeal and replace’ was once a unifier for the GOP. Now it’s an albatross.	https://www.washingtonpost.com/politics/repeal-and-replace-was-once-a-unifier-for-the-gop-now-its-an-albatross/2017/06/27/3cce8eaa-5b72-11e7-a9f6-7c3296387341_story.html?hpid=hp_hp-top-table-main_take-627pm%3Ahomepage%2Fstory&utm_term=.c61ca29ca106
Vote Delayed as G.O.P. Struggles to Marshal Support for Health Care Bill	https://www.nytimes.com/2017/06/27/us/politics/republicans-struggle-to-marshal-votes-for-health-care-bill.html