

# ライトウエイト・メタデータの 応用事例とその可能性

データセクション(株)

橋本大也

2005/7/19@人工知能学会

セマンティックウェブとオントロジー研究会

# 内容

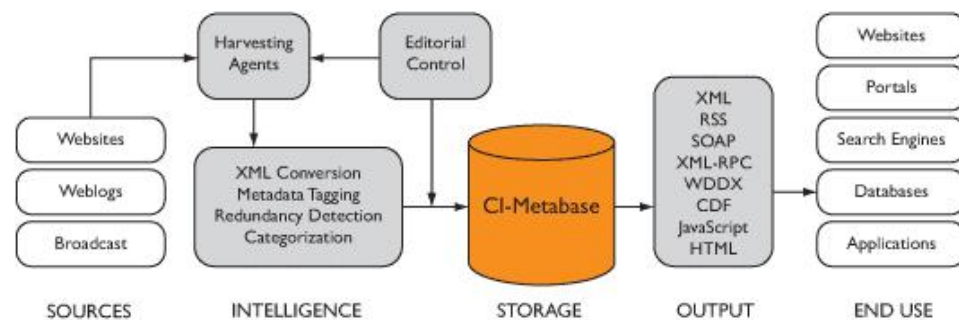
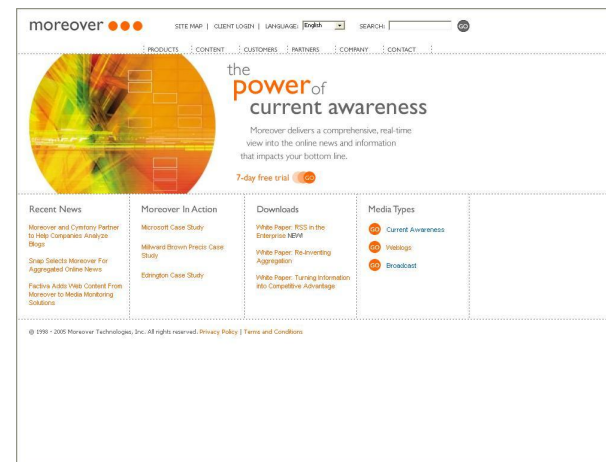
1. ライトウエイト・メタデータの応用例として、いくつかのWebサイトおよびサービスを取り上げ
2. 現状のWeb技術について概説する。
3. また、これらの展望や可能性について議論する。

# I メタデータ応用事例

- 以下の定番を除外しています
- ブログ検索
  - テクノラティ、未来検索Livedoor、はてな
  - など一般的なブログ検索
- ブログリーダー
  - Bloglines、Glucose
  - など一般的なRSSリーダー
- 一般的なソーシャルネットワーク、ブックマーク
- 今日の参加者にとってCurrentでEmergingだと思われる事例を10件

# 事例1 アグリゲーター Moreover.com

- Moreover.com
  - <http://www.moreover.com>
- 1万件のRSS情報を集約
- 125カ国、26言語
- 15分間隔で更新、380のカテゴリに分類
- 人間の目で選んだRSS
- 1日14万件の最新ニュース
- (1)Current Awareness、(2)Weblogs、(3)Broadcastの3分類
- カスタマイズビジネス展開
  - 顧客にマイクロソフト他

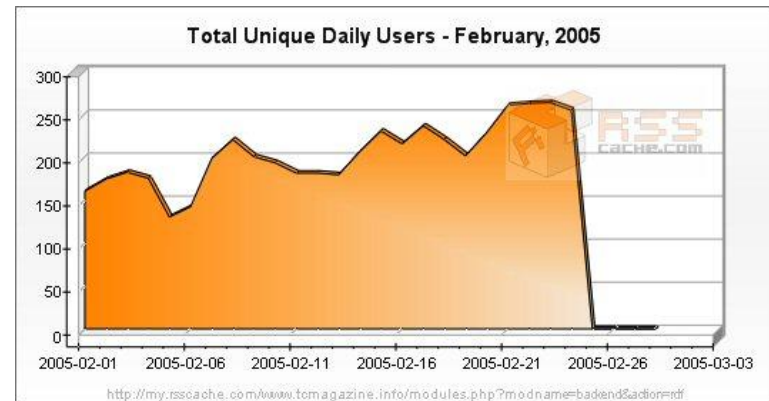


参考: 検索のテクノラティ  
<http://technorati.com/>

# 事例2 プロクシーサーバ RSSCache.com

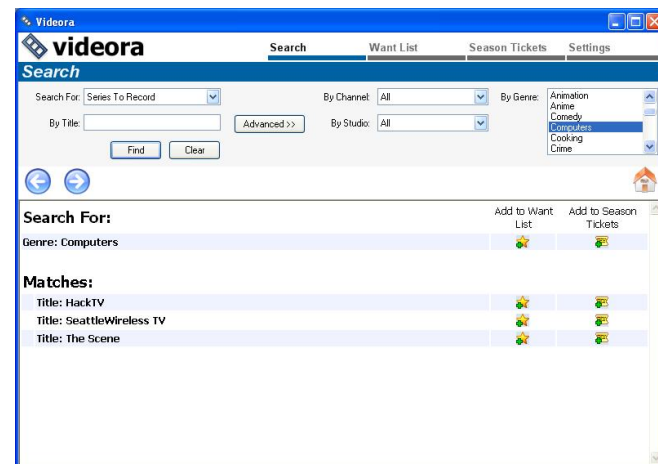
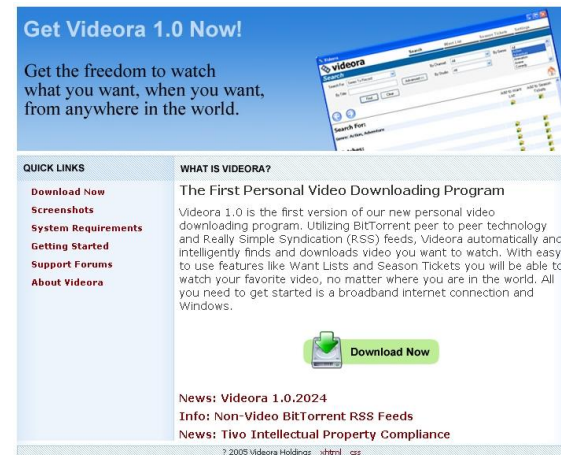
- RSSの負荷分散のためのキャッシュ機能を提供する
- RSSのURLからキャッシュ機能のURLを生成する  
(例)
  - <http://my.rsscache.com/www.cacert.org/rss.php>
- 登録RSSのアクセス統計データを表示できる
- RSSの人気ランキング情報やRSS広告の提供

The screenshot shows the RSSCache.com website. At the top, it says "The bandwidth saver solution for your RSS feeds!". Below this, there are sections for "What is RSSCache.com:", "How it works!", "WEBMASTERS", "USERS", and "ENTERPRISE SOLUTIONS". The "WEBMASTERS" section includes a code snippet: `http://my.rsscache.com/www.your.site.com/feed.xml`. The "USERS" section mentions "Generate URL". The "ENTERPRISE SOLUTIONS" section mentions "Learn more about our solution based on the Microsoft .Net technology!".



# 事例3 RSS+P2P Videora

- Videora
  - <http://www.videora.com/>
- P2Pファイル共有のBittorrent上のファイル情報をRSSとして取得
- キーワードにマッチするビデオファイルを自動ダウンロードする
- RSS + Share



# 事例4 雑誌の記事をRSS配信 FindArticles.com

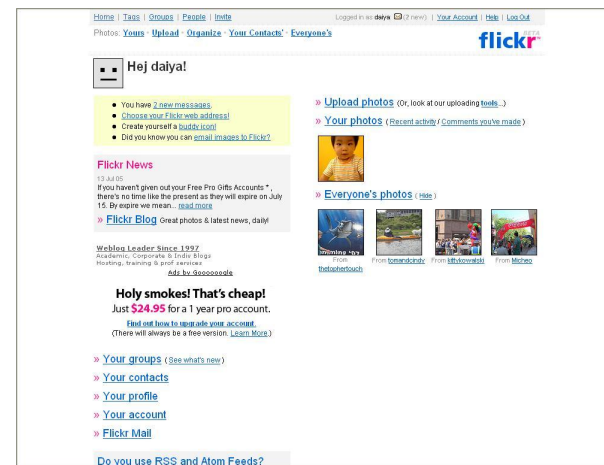
- FindArticles
  - <http://www.findarticles.com/>
- 数千の雑誌記事の記事情報(本文含む)を1984年まで遡ってデータベース化
- 収録数1000万記事以上
- RSSとして最新情報と検索結果を提供
- 検索結果から有料記事を購読することができる



- 参考: 米国雑誌の3分の1がRSS公開  
<http://publications.mediapost.com/index.cfm?fuseaction=Articles.san&s=31662&Nid=14155&p=276816>

# 事例5 画像のメタデータ Flickr!と派生サービス

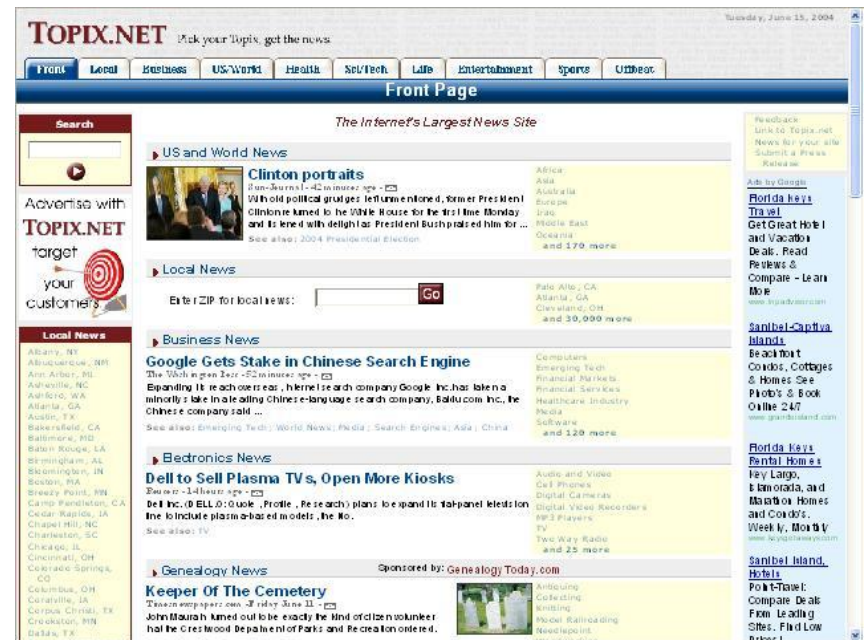
- Welcome to Flickr!
  - <http://www.flickr.com/>
  - 写真にキーワード(タグ)をつけてアップロードし、ユーザ間で共有するサービス
- API応用
  - Mappr
    - <http://mappr.com/>
  - Flickr Related Tag Browser
    - [http://www.airtightinteractive.com/projects/related\\_tag\\_browser/](http://www.airtightinteractive.com/projects/related_tag_browser/)
  - flickr graph - marcos weskamp
    - <http://www.marumushi.com/apps/flickrgraph/>
  - Flickr/TiVo
    - [http://home.comcast.net/~major\\_clanger/TiVo/](http://home.comcast.net/~major_clanger/TiVo/)





# 事例6 メタデータ自動生成ニュース Topix.net

- Topix.net
  - <http://www.topix.net/>
- ニュースサイトのメタデータを自動分類、重要度判断しトピック別の新聞を自動作成
- 1万のソースサイト、30万のトピック(3万の都市、5500の企業、4.8万人の著名人、1500のスポーツを含む)



参考: GoogleNews

<http://news.google.co.jp/nwshp?hl=ja&ned=jp>

# 事例7 リアルで書籍メタデータ BookCrossing

- BookCrossing
  - <http://www.bookcrossing.com/>
- Webでプレートを印刷
- 書籍にプレートを貼り付けIDを書いて街に“放流”する
- 本を拾った人はWebで感想を書いて再放流
- 街が図書館となり書籍の歴代保有者が感想データを蓄積、ユーザ同士で交流
- 世界で37万人、220万冊が無料で流通している実績
- 参考 PlateMatch
  - <http://www.platemark.com/>



so many books, so little time  
books registered: 2,280,454  
total books available

bookcrossing.com

Welcome to BookCrossing!

Did you catch a BookCrossing book?  
Woo hoo! Please make a journal entry right away, using its BCID number (look inside the front cover):

BCID:  Goal:

What is BookCrossing?  
bookcrossing n. the practice of leaving a book in a public place to be picked up and read by others, who then do likewise.  
(added to the *Cambridge Online English Dictionary* in August 2009)

You've come to a friendly place, and we welcome you to our book-lovers' community. Our members love books enough to let them go—into the wild—to be found by others.

Howdy!  
Hola!  
Ciao!  
Bonjour!  
Guten Tag!

I'm a *very special* book.  
You see, I'm traveling around the world, making new friends wherever I go. I hope I've found another friend in you!

Visit [www.bookcrossing.com](http://www.bookcrossing.com) and write a brief journal entry with my BCID number (below). Then keep my dream alive:  
**READ and RELEASE ME!**

BCID:

First Registered By:

When & Where:

**bookcrossing.com**  
it's the karma of literature, free of cost and absolutely anonymous. Make a journal entry on this book today for a chance to win a \$100 book shopping spree!

© 2001-2004 bookcrossing.com

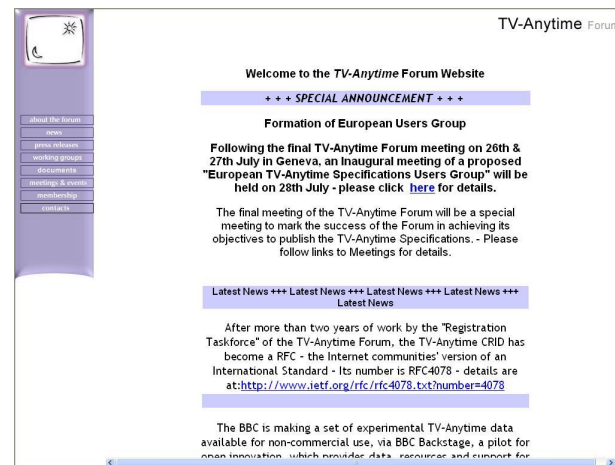
# 事例8 テレビのメタデータ テレビブログ、TV-Anytime

- テレビブログ
  - <http://www.tvblog.jp>
  - テレビ番組情報に対してトラックバックを受け付けるブログASPサービス
  - ブログのコメントRSS
  - 動画メタデータも視野に開発
- TV-Anytime
  - <http://www.tv-anytime.org/>
  - テレビ番組情報の国際標準規格。英国BBCなどが採用している。
  - <http://backstage.bbc.co.uk/feedstvradio/>



The screenshot shows the TVblog website interface. At the top, there's a navigation bar with links like 'ブログを始める' and '会員登録ページ'. Below that is a calendar for July 2005, with the current date 2005/07/11 highlighted. A table below the calendar lists TV programs for the current day, categorized by channel and time slot. The table has columns for channel (e.g., NHK総合, NHK教育), program name, and other details.

チャンネル	番組名	放送時間
NHK総合	スポンサーごきす	7:00 - 7:00
NHK総合	きよひの出来事	7:00 - 12:00
NHK総合	NANNI+キュメント	12:00 - 19:00
NHK総合	ニュース、気象情報	19:00 - 24:00
NHK教育	芸術花舞台	
日本テレビ	JNN(N)0	
TBSテレビ	Jスポーツ0	
フジテレビ	2005F1イギリスGP(決勝0)	
テレビ朝日	やべっちFC0	
テレビ東京	SHOWBIZ CO UNTDOWN0	
	Get-Sports0	
	TXND00	
	知る明練0	



The screenshot shows the TV-Anytime Forum website. It features a navigation menu on the left with links like 'about the forum', 'news', 'press releases', 'meeting notices', 'discipline', 'meeting & social', 'membership', and 'contact'. The main content area has a header 'Welcome to the TV-Anytime Forum Website' and a 'SPECIAL ANNOUNCEMENT' section. The announcement discusses the formation of an European Users Group and the final meeting of the TV-Anytime Forum in Geneva. It also mentions the registration of the TV-Anytime CRID as an International Standard (RFC-4078).

TV-Anytime Forum

Welcome to the TV-Anytime Forum Website

+++ SPECIAL ANNOUNCEMENT +++

Formation of European Users Group

Following the final TV-Anytime Forum meeting on 26th & 27th July in Geneva, an Inaugural meeting of a proposed "European TV-Anytime Specifications Users Group" will be held on 28th July - please click [here](#) for details.

The final meeting of the TV-Anytime Forum will be a special meeting to mark the success of the Forum in achieving its objectives to publish the TV-Anytime Specifications. - Please follow links to Meetings for details.

Latest News +++ Latest News +++ Latest News +++ Latest News +++ Latest News

After more than two years of work by the "Registration Taskforce" of the TV-Anytime Forum, the TV-Anytime CRID has become a RFC - the Internet communities' version of an International Standard - its number is RFC-4078 - details are at: <http://www.iETF.org/rfc/rfc4078.txt?number=4078>

The BBC is making a set of experimental TV-Anytime data available for non-commercial use, via BBC Backstage, a pilot for open innovation, which provides data, resources and support for

# 事例9 目標と達成のオントロジー 43Things






- 43 Things
  - <http://www.43things.com/>
- 達成したい目標(本を書く、10キロ痩せる、etc)を登録する
- 同じ目標を持つユーザのブログを一覧、検索できる
- ソーシャルブックマーク、共有キーワード、ランキング表示機能などを提供する

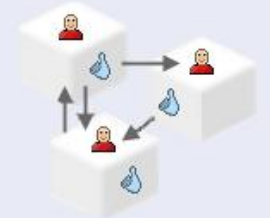
The screenshot shows the 43Things website interface. At the top, there is a navigation bar with links for Home, Zeitgeist, Your 0 Things, Log In, and a search bar. The main heading is "43 Things". Below this, a question "What do you want to do with your life?" is followed by a search input field and a "I want to do this" button. The main content area displays a list of goals shared by 43,373 people in 4,854 cities, including "make people smile", "be more positive", "become a firefighter", "be happy with who I am", "ride critical mass", "read a book a week", "Ankur Gupta wants to make a collage", "Have fun", "minimize my material possessions", "laugh every day", "Rule a small island nation", "sail around the world", "Nikki wants to stop biting my nails", "be less judgemental", "be enlightened", "write love letters", "learn to sew", "lose weight", "overcome depression and anxiety", "love and be loved in return", "learn to SCUBA dive", "walk my dog everyday", "become more artistic", "actually learn to play my guitar", "have normal sleep hours", "own a ranch", "queenbitch wants to get pool glasses", "balance my hormones so PMS stops being a problem", "Lose Weight and Get Healthy", "learn first aid", "buy a new car", "learn areek", "clean my house", "be more politically active", "Stop screwing around on the internet", "be physically fit", "Iesthestar wants to see nine inch nails live", "play guitar", "Build a computer", "learn to dance flamenco", "go to Antarctica", "fix my ipod", "graduate college with a bachelors degree", "go back to yellowstone", "keep a sketchbook", "meetup with other 43 things people in Ann Arbor, Michigan", "visit california", "Go to the dentist", "Tithel learn to let go", "buy my own house", "meet new friends", "work in barcelona", "John Resiq wants to Get (or renew) a passport", "be fluent in Japanese", "Shuffles52 wants to ride an elephant", "Learn Flash", "travel all over the world", "sing in a choir again", "Read more often", "Become a better programmer", "Graduate", "listen more", "acquire an island", "go to the Farmer's Market more often", "Attend every tennis grand slam event", "loopy1 wants to take my dog for a walk more often", "write a novel", "publish my poetry", "start exercising", "Go to the Edinburgh Fringe Festival", "Be nicer to my husband", "hike the appalachian trail", "Visit Ireland", "go to the gym regularly". On the right side, there are sections for "Discover what's important, make it happen, share your progress. Find your 43 things. Learn more--", "Top Cities" (listing cities like Seattle, Chicago, London, etc.), and "Today's Tags" (listing tags like travel, health, music, etc.).

# 事例10 ソーシャルネットワーク検索 StumbleUpon

- StumbleUpon
  - <http://www.stumbleupon.com/>
  - ツールバーで表示中のWebをユーザ評価する
  - 専用ブログと連動
  - 25万ユーザ
  - 200万サイトに対して1億超の評価情報
  - 500のトピックで分類
  - 登録サイトの検索や人気ランキング
- 参考: Eurekaster
  - <http://eurekaster.com/>



StumbleUpon uses   ratings to form **collaborative opinions** on website quality. When you stumble, you will only see pages which friends and like-minded **stumblers** () have liked. Unlike search engines or static directories, this allows for a true "democracy of the web" – all SU members have a say ( or ) as to whether a page should be passed on.



# 事例11 音楽情報の自動分類

## Moodlogic.com, MusicBrainz!

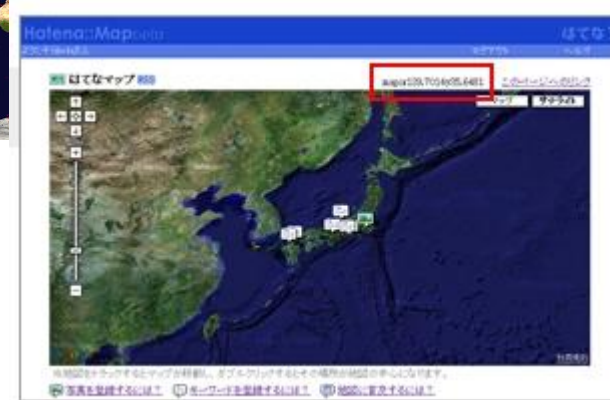
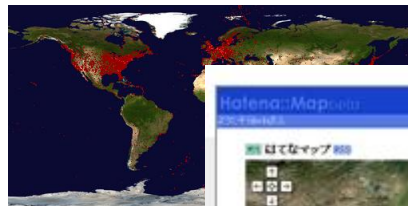
- Moodlogic.com
- <http://www.moodlogic.com/>
- 楽曲についての感性データを手作業で作成しデータベース化。
- MP3再生管理アプリケーションを配布。サーバからメタデータを配信。
- ユーザPCのMP3ファイルを感性データで検索することを可能にする。
- ロマンチックな80年代の曲、ハッピー、アグレッシブ、アップビートなど感性語、ブルースやカントリー、クラシックロックなどジャンル語、演奏テンポなどで検索することが可能。
- MusicBrainz!
- <http://musicbrainz.org/>



参考: MusicID、CDDb  
<http://www.cddb.com/>

# 事例12 地域情報とSNS、Blog InsiderPages, GeoURL、はてなマップ

- InsiderPages
  - <http://www.insiderpages.com/>
- The Yellow Pages written by friendsがコンセプト。ソーシャルネットワーク上で自分の良く知っている店舗などの地域情報を登録する。友人関係上で信頼できる地域情報のみを交換する仕組み。
- GeoURL (2.0)
  - <http://geourl.org/>
  - RSSに緯度経度を記述することで地図上にマップ
- はてなマップ
  - <http://map.hatena.ne.jp/>
  - 地図へトラックバック



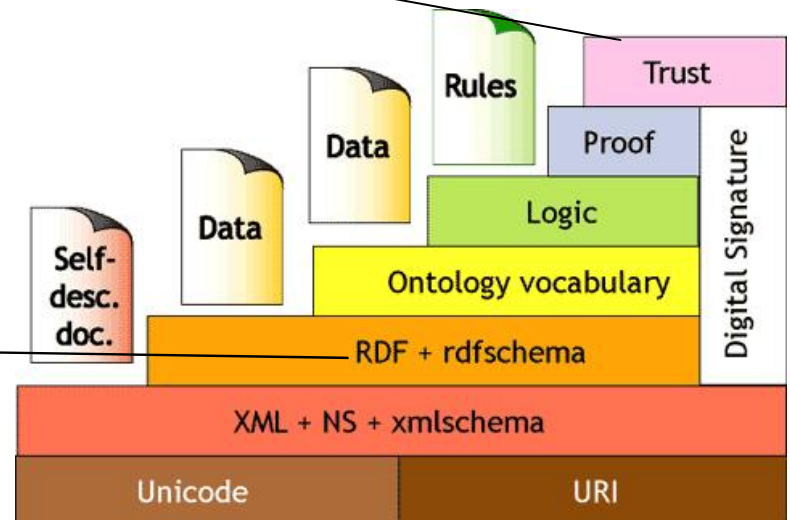
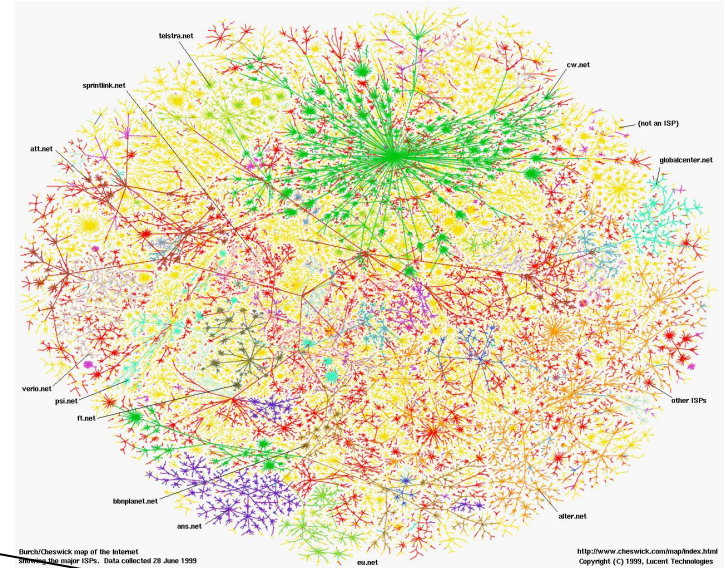
## Ⅱ 事例の概説

- 個別に便利だが、セマンティックWeb的でないサービスが多い
- 草の根インデクサーとして機能している
- “マッピング”が有意味に行われていない
- Webの情報全体を統合、高次化するような方向性が見えてこない
- 今日のテーマは「ライトウェイト」？
  
- そもそも...



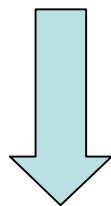
# なぜ今、メタデータが必要なのか？

- インターネット上のファイルの量が爆発的に増大して必要な情報を見つけることが難しくなった。
- 字面だけでなく意味で整理することで信頼(Trust)できる情報を探し出せるようにしよう(次世代のWeb、セマンティックWeb)。
- 情報のエントロピー(乱雑度)を下げる技術群が必要だ(実現ピラミッド)。
- メタデータがその基盤になる

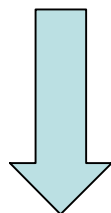


# メタデータで整理するとは

何億枚の文書が乱雑に置かれている状況  
→ どこに何があるのか分からない

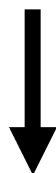


メタデータというカードを作成する  
→ 著者、見出し、要約、作成日、棚番号などをカード化

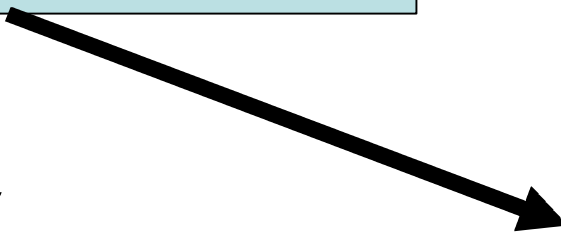
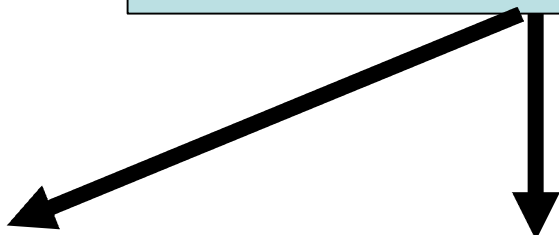


カードで索引が整理され文書が一覧、検索可能になった状況  
→ すぐに欲しい情報が見つかる

# 整理されるとどうなる？



情報流通の効率化  
メタデータ(カード)なら流通できる



情報の発信者と読み手  
が会う

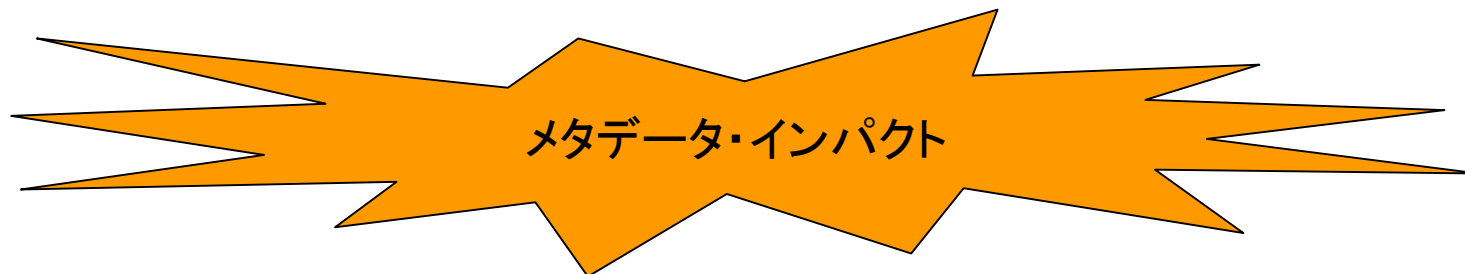
コミュニティの活性化

モノの売り手と買い手  
が会う

マーケティングの最適化

情報流通全体の  
見通しがよくなる

情報技術の高次化



メタデータ・インパクト

# 流行したが効率化と高次化が進まない メタデータ

- RSSがメタデータとしてではなく、配信メディアとして流行した結果、データ化して使われていることが原因(例:RSS広告) “ベタ”データ問題
- メタデータのメタ性を再考する必要あり
- オントロジーマッピングによる高次化に必要なTrustが不足しているためセマンティックWebとして高次化できない現状(例:ブログのカテゴリ、スパム)
- 生成方法として大きく2種類のメタデータが流通している

# Ⅲ 展望と可能性

# 集計系と編集系のメタデータ

## • 集計系

- 機械が自動的に付与
  - 新着更新情報
  - 人気ランキング
  - 検索結果
  - 経済指標などの数値
  - 公的大本営発表
- 計算可能なデータ

## • 編集系

- 人間が手作業で付与
  - 重要度、緊急度など
  - ニュースバリュー
  - 高度なカテゴリ分類
  - 論評、意見、批評
- 計算不可能なデータ



技術でできる範囲が拡大している

セマンティックWeb、人工知能、言語処理

# メタデータは”ネタ”データへ

- 骨格となるストリクトなメタデータに、コミュニティがラフなデータを付加してリッチにするモデル (Community Generated Metadata)
  - Ex. Wikipedia, ODP
- StrictSemantics  $\Leftrightarrow$  RoughSemantics (大向)
- RoughSemanticsの集まるネタノードとしてのStrictSemanticsが求められている
- Trustに対してStrictな集計系、Roughな編集系

# 電車男というRoughメタデータ

- 「電車男」、「今週妻が浮気します」...ネタが尽きないCGMのトレンド
- ネタデータ=コミュニケーション揮発性の高いメタデータ
- ネタデータをベースにコメントやトラックバックで情報(アノテーション)を集めることは容易に
- Invisible&Deepなネタデータに大きな可能性
- 信頼性は？

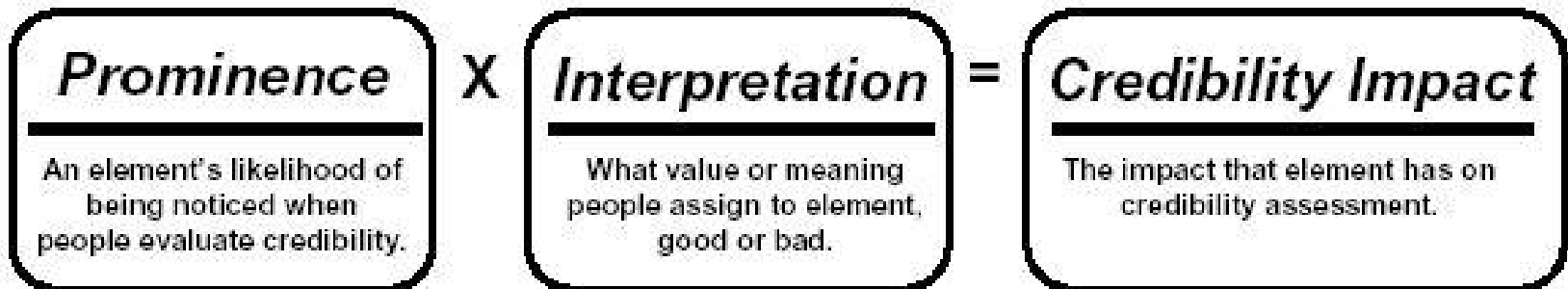


# 参考：メタデータの抱える問題点

- Metacrap
  - <http://www.well.com/~doctorow/metacrap.htm>
- 2.1 People lie 人は愚かである
- 2.2 People are lazy 人は怠け者である
- 2.3 People are stupid 人はお馬鹿である
- 2.4 Mission: Impossible -- know thyself そもそも不可能である
- 2.5 Schemas aren't neutral 非中立的である
- 2.6 Metrics influence results
- 2.7 There's more than one way to describe something 同じことをいくつもの方法で表現できる

# 参考：信頼性とPI理論

- Prominence-Interpretation Theory:  
ユーザはどうやってオンラインで信頼性を評価するのか（Stanford Persuasive Technology Lab）  
<http://credibility.stanford.edu/pdf/PITheory.pdf>
- 1 ユーザは目立つものを見つける（際立っているという評価度）→ 集計系メタデータ
  - 2 ユーザはそれを解釈する（解釈による評価度）  
→ 編集系メタデータ



# 今後の課題

- セマンティックWebのTrustの価値をどこから発生させるか
  - 信頼できる機関のStrictなメタデータを増やす
  - Roughなメタデータを処理して信頼性を高める
- 次のステップ“マッピング”に耐えうる信頼度の高いメタデータを統合するサービスが次世代
- Webサービスというアクションとの統合で「セマンティックWebエージェント」の時代へ

# ありがとうございました

- 「メタデータ、ベタデータ、ネタデータ」
- Strict & Rough Semantics
- Trust
  
- 本日のファイルを下記にアップロードします
- 情報考学 Passion For The Future
  - <http://www.ringolab.com/note/daiya/>
  - またはGoogleで「橋本大也」