

メーリングリストのメタ情報の RSS/iCalendarによる配布

神島 敏弘 (産業技術総合研究所)

<http://www.kamishima.net/>

第21回 人工知能学会 全国大会 (2007.6.20-22)

はじめに

メーリングリストからメタ情報を抽出し、RSSやiCalendarとの形式で配信するシステムの実装

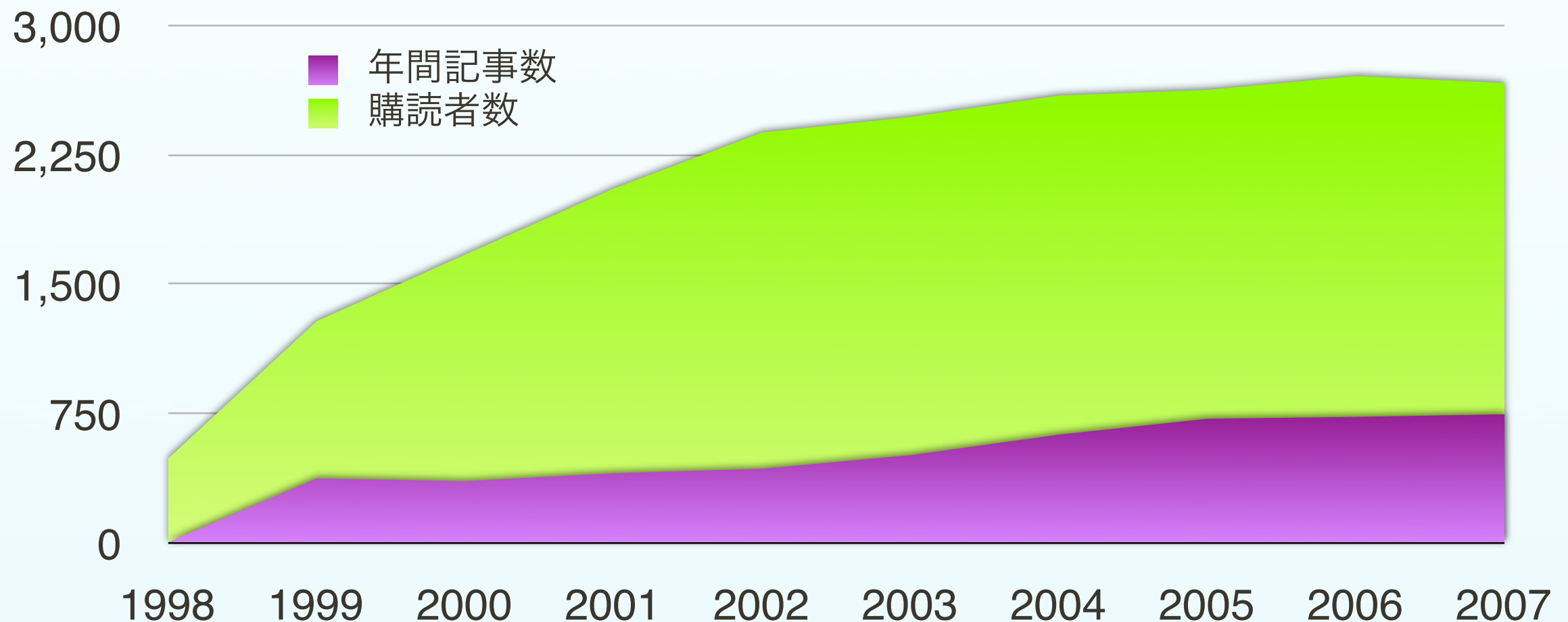
人工知能学会からのお知らせメーリングリスト

以後 **AI学会ML**

- ⌚ 人工知能に関連するアナウンス情報を配信
- ⌚ 1998年6月1日より運用を開始
- ⌚ moderatedな運用体制
- ⌚ Subjectに開催・締切日情報を付加
- ⌚ 2006年6月1日よりRSS/iCalendar形式による配布を開始

AI学会ML：現状

- 9年間運営
- 最近の購読者数は 2,600～2,700人の間で推移
- 配信記事数は単調に増加，9年目の記事数は 741件 (1日 2.03件)
- 多数の記事を効率的に処理する方法が必要



AI学会ML：ヘッダ情報

ヘッダ情報：イベントや論文募集の開催日や締切日の情報を，管理者が人手でメールのSubject行に手作業で付加

記事番号

AI学会MLの
記事の通し番号

記事の種類

MEETING=イベント情報
CFP=原稿募集

付属情報

h=開催日
d=締切日

Subject: [jsai-ann 4738] CFP;h=070623;d=070417:
MYCOM2007(第8回AI 若手の集い)

略タイトル

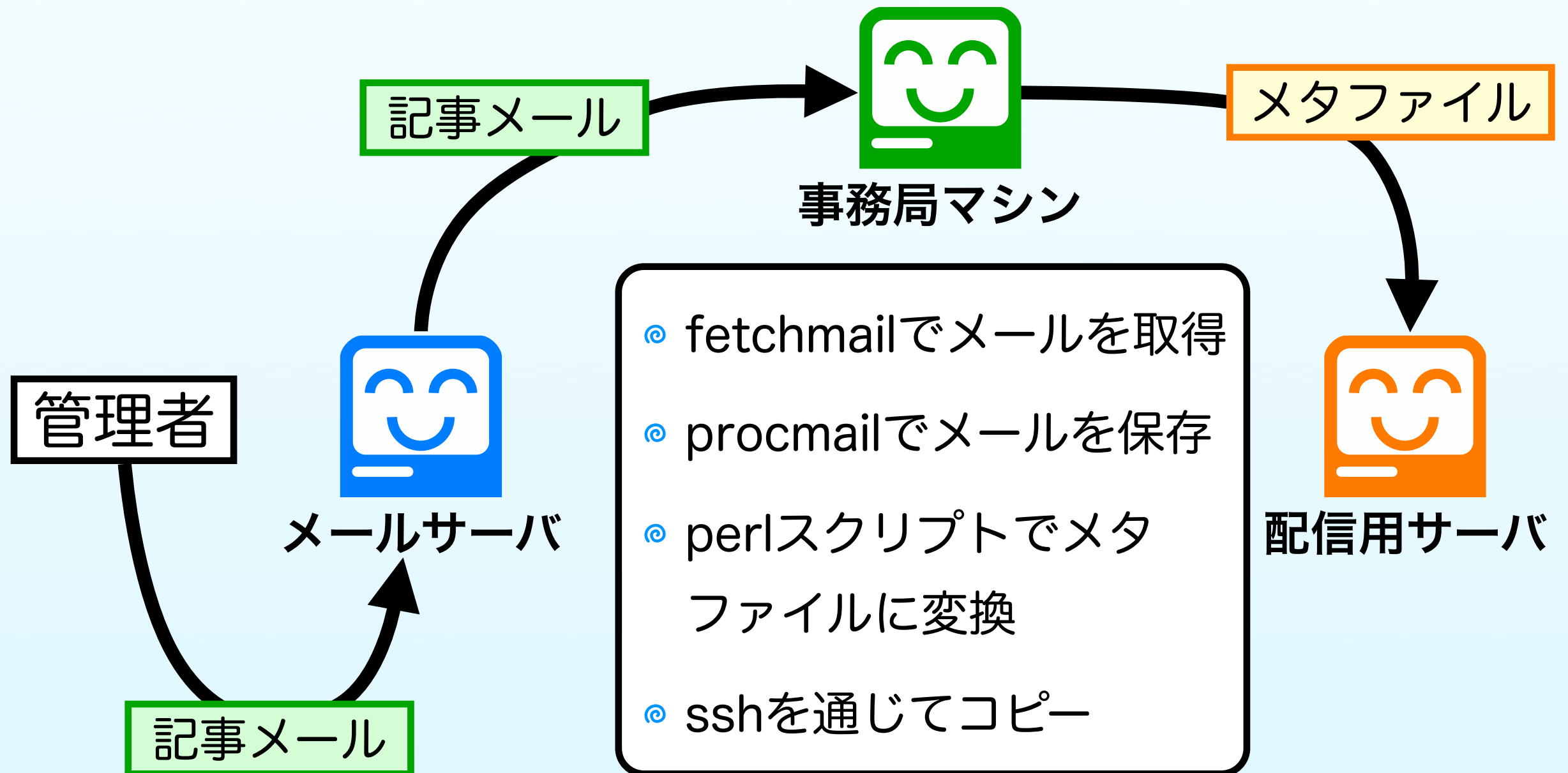
イベントの省略名

詳細タイトル

会議の正式なフルネーム

システム構成

学会によって運営するため、常駐の管理体制は準備できない
システム構成はできるだけ簡素に



RSS (RDF Site Summary)

RSS：ブログやWebの更新情報をpull型で配信するためのメタ情報フォーマット

- ◎ 幾つかのフォーマットが並立
- ◎ セマンティックWebのRDF (Resource Description Framework) に準拠したRSS1.0(RDF Site Summary)を採用
- ◎ 最近の記事, 開催日情報, 締切日情報の3種類のファイルを配布



MacOSXのWebブラウザ
SafariによるRSS表示

RSSの語彙

RSSの基本語彙

- channel部：全体のタイトル, URL, 説明
- item部：個々の記事のタイトル, URL, 説明

Dublin Coreの語彙

- 全体の発信日時, 作成者, 著作権表示.
- 個々の記事の発信日時

RSSの拡張語彙

- Syndication：更新周期などの同期情報
- content：HTMLで記述した記事の一部
- Event：イベントの開催日時, 種類, 主催者, 開催場所

```
1 <?xml version="1.0" encoding="utf-8" ?>
2 <?xml-stylesheet type="text/xsl" href="http://www
...
3 <rdf:RDF
4   xmlns="http://purl.org/rss/1.0/"
5   《ファイル中で利用する語彙に関する宣言》
6   xml:lang="ja">
7 <channel rdf:about="http://www.ai-gakkai.or.jp/rs
...
8 <dc:language>ja</dc:language>
9 <title>人工知能学会 ML(200)</title>
10 <link>http://www.ai-gakkai.or.jp/jsai/ml/</link>
11 <description>《このファイルの案内文》</description>
12 <dc:date>2007-04-11T16:02d:02d+09:00</dc:date>
13 <dc:rights>Copyright(c), The Japanese Society for
...
14 <dc:publisher>人工知能学会</dc:publisher>
15 <dc:creator>人工知能学会</dc:creator>
16 <sy:updatePeriod>hourly</sy:updatePeriod>
17 <sy:updateFrequency>2</sy:updateFrequency>
18 <items>
19 <rdf:Seq>
20 <rdf:li rdf:resource="http://jsai-ann:ai-gakkai@...
  《...中略...》
21 <rdf:li rdf:resource="http://jsai-ann:ai-gakkai@...
22 </rdf:Seq>
23 </items>
24 </channel>
25 <item rdf:about="http://jsai-ann:ai-gakkai@mlwww...
26 <title>C&amp;0:RR-2007</title>
27 <link>http://jsai-ann:ai-gakkai@mlwww.iijnet.or...
28 <description>《記事のテキストによる説明》</description>
29 <content:encoded>《記事の HTML による説明》</content:...
30 <dc:date>2007-04-11T08:10:32+09:00</dc:date>
31 <ev:startdate>2007-08-21</ev:startdate>
32 </item>
33 <item rdf:about="http://jsai-ann:ai-gakkai@mlwww...
34 <title>IPC-07</title>
  《...中略...》
35 <ev:startdate>2007-06-18</ev:startdate>
36 </item>
37 </rdf:RDF>
```

イベント開始日の問題

タグの表す「意味」に配慮し、情報の再利用を可能に

itemタグが表すもの「記事」or「イベント」？

itemが示す記事が示すイベントの開催日時だとすると……

ev:startdateの記述

```
<item rdf:about="http://...">  
<dc:date>2005-07-26T17:49:20+09:00</dc:date>  
<foaf:topic ev:startdate="2006-02-13"/>  
</item>
```

foaf:topicを使って記事の内容の日付であることを示す



Eventモジュールのドラフトの例では……

ev:startdateの記述

```
<item rdf:about="...">  
<ev:startdate>2001-09-18</ev:startdate>  
</item>
```

採用

ev:startdateが「記事の内容の日付」という解釈

RSSのその他の記述

item の title タグ

- ◉ 開始日や締切日情報は ev:startdate のタグに含めた
- ◉ ev:startdate をサポートしているRSSリーダは少ない
- ◉ item の title タグの先頭に日付を書き込んで、利用者から参照できるようにした

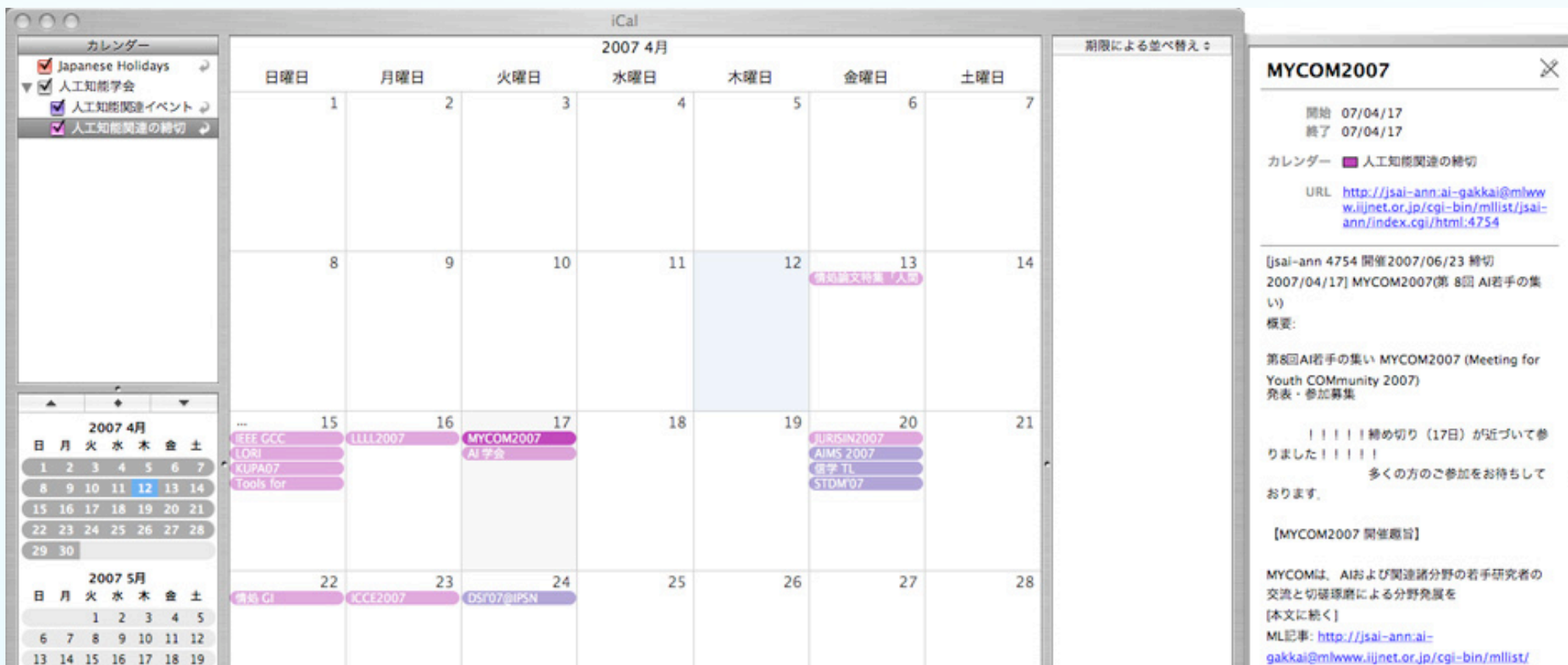
item の description と content:encoded タグ

- ◉ メタ情報：記事番号と開催日・締切日の情報
- ◉ 記事タイトル：詳細タイトルも含めた記事のタイトル
- ◉ 関連WWW：記事中から抜粋したURL
- ◉ 記事の抜粋：記事の先頭20行

iCalendar

- ◉ iCalendarはスケジュール情報を相互運用するためのフォーマット
- ◉ RFC2445で標準化
- ◉ 前身として vCalendar がある
- ◉ iCalendarで表された情報を交換するプロトコル calDAV もある
- ◉ AI学会MLでは、イベント開催日と論文募集締切日の2種類を配布

MacOSXのスケジュール管理ソフトiCalによる表示



iCalendar : ファイル構造

VCALENDAR部



- ◉ カレンダーファイル全体の情報：タイトル, 暦の種類, 説明, 配信方法, タイムゾーン

TIMEZONE部



- ◉ タイムゾーンの实体. 年代による違いも扱う

VEVENT部



- ◉ DTSTAMP : 配信日時
- ◉ DTSTART, DTEND : 開始・終了日時
0時開始で24時終了なら1日中のイベント
- ◉ SUMMARY : タイトル, 略タイトルを表示
- ◉ DESCRIPTION : イベントの説明
- ◉ URL : イベントのURL

```
1 BEGIN:VCALENDAR
2 VERSION:2.0
3 PRODID:-//www.ai-gakkai.or.jp/ics/event 1.0//EN
4 CALSCALE:GREGORIAN
5 X-WR-CALNAME:人工知能関連イベント
6 X-WR-CALDESC:《このファイルの案内文》
7 METHOD:PUBLISH
8 X-WR-TIMEZONE:Asia/Tokyo
```

```
9 BEGIN:VTIMEZONE
10 TZID:Asia/Japan
11 LAST-MODIFIED;VALUE=DATE:20070411
12 BEGIN:STANDARD
13 DTSTART:19510908T020000
14 TZOFFSETTO:+0900
15 TZOFFSETFROM:+1000
16 TZNAME:JST
17 END:STANDARD
18 END:VTIMEZONE
19 BEGIN:VEVENT
20 DTSTAMP;TZID=Asia/Japan:20070410T174017
21 DTSTART;VALUE=DATE:20070623
22 DTEND;VALUE=DATE:20070624
23 SUMMARY:MYCOM2007
24 UID:jsai-ann-4754@ai-gakkai.or.jp
25 URL;VALUE=URI:http://jsai-ann:ai-gakkai@mlwww.ii...
26 DESCRIPTION:[jsai-ann 4754 開催 2007/06/23 締切 200...
27 I 若手の集い) \n 概要:\n\n 第 8 回 AI 若手の集い MYCOM...
28 /mllist/jsai-ann/index.cgi/html:4738\n
29 END:VEVENT
30 END:VCALENDAR
```

レコードリンケージ

レコードリンケージ：現実世界で同じ実体を指し示す，データ中のエントリーをまとめること

別名：identity uncertainty, entity resolution, reference matching

同じイベントを指し示す複数の記事が発生する理由

- ◉ 同じイベントについて，異なる内容の記事が配布される
例：論文募集と参加案内，本会議とワークショップの募集
- ◉ 期日前になるとReminderとして同じ記事が配布される

なぜ難しいのか？

- ◉ 表記にばらつき：JSAI2007, JSAI'07, 第21回 AI学会 全国大会
- ◉ コンテキスト依存性：英語の記事と日本語の記事は同じ？
ワークショップの募集と本会議の募集は同じ？

ヒューリスティックな解決

採用した方法：ヒューリスティックな書き換え規則で，省略タイトルを正規化．同じ名前のエントリーを，最も新しいエントリーに併合

- ◉ 2桁表記('07や-07)の年が，開催日の年と一致するなら4桁に
- ◉ “SIG-”の省略や，代表的な学会名の正規化(人工知能学会⇒AI学会)
- ◉ 空白や記述記号の削除と，英字大文字への変換

違うコンテキストの同じ名前：本会議募集とワークショップ募集などは，会議名が同じなので，併合されてしまう．

- ◉ たとえ過剰に併合されてしまっても，完全には抹消せずに，利用者がアクセスできる手段を残す．DESCRIPTION内に，関連記事としてリストアップしておく．

機械学習を用いた解決

もし本格的にレコードリンケージ問題を扱うなら……

クラス分類問題

- ⦿ 同じ実体を表すエントリー対を正例，違う実体を表す対を負例としてクラス分類問題として解く
- ⦿ 同じ実体になる度合いが，しきい値以上の対をまとめることで，レコードリンケージを行う

半教師ありクラスタリング

- ⦿ 同じ実体を表すエントリー間を同じクラスタにならなくてはならないmustリンク制約で，そうでないエントリー間を違うクラスタであるcannotリンク制約で結び，これらの制約を考慮してクラスタリングする

デモ

おわりに

まとめ

- ◉ AI学会MLの記事からメタ情報を抽出し、それをRSSとiCalendar形式で配布
- ◉ その設計と実装についてまとめた

人工知能学会 RSS/iCalendar ファイル

<http://www.ai-gakkai.or.jp/jsai/event/>

おまけ

- ◉ 「朱鷺の杜Wiki」 機械学習やデータマイニングについてのWiki

<http://ibisforest.org/index.php?FrontPage>