General Metadata Hospital
Part II: Chronological Metadata

Ying Wang Xiangrui Meng

Institute for Computational and Mathematical Engineering
Stanford University

CADS 2009

Outline

Outline

Desired Service

Enable browsing by specifying a date or a date range on a timeline.

Summer Tasks

Tasks:

- Make date format in metadata consistent.
- Generate or improve date and date range from available information in the record or text, using information in the metadata record, chronological references and named events extracted from the text, and external sources.

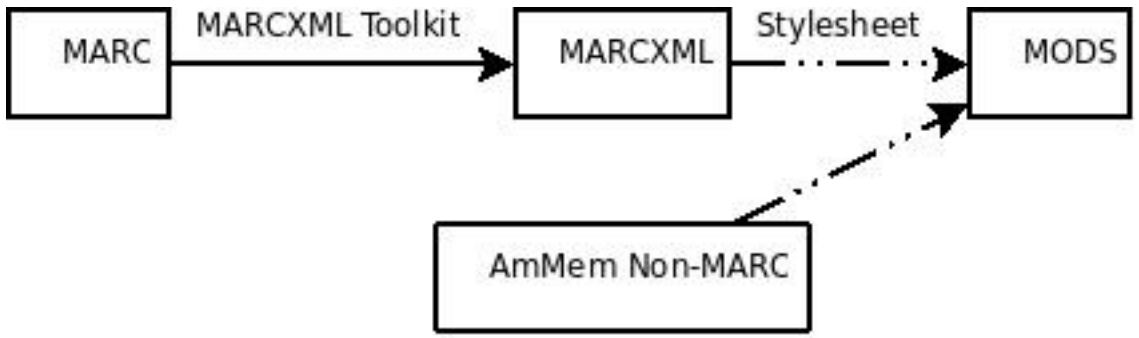
Data:

- Five MARC metadata files, including Civil War Photographs and America at Work.
- Five Non-MARC metadata files from the American Memory collections, including Abraham Lincoln Papers and Frederick Douglas Papers.

Resource Description Formats

- MARC: MAchine-Readable Cataloging.
- Non-MARC: An XML schema for AmMem metadata files.
- MODS: Metadata Object Description Schema. MODS is a derivative of MARC, richer than Dublin Core but much simpler than MARC.

The metadata records we received are in either MARC or Non-MARC format. Our target format for remediated metadata is MODS. LC provides conversion toolkits and stylesheets to transform MARC and Non-MARC records into MODS records.



Chronological Metadata Elements

Format	MARC	Non-MARC	MODS
Fixed length	008/00-05		<recordCreationDate>
	008/07-14		<dateIssued>
	008/11-14 if 008/06=t		<copyrightDate>
Textual		record_create_date	<recordCreationDate>
	260$c	publication_date	<dateIssued>
	260$g	date_sorter, text_date	<dateCreated>
	033$a		<dateCaptured>
	046$m,n		<dateValid>
	046$j		<dateModified>
	X00$d		<name><namePart type="date">
	X10$d,f		<name><namePart type="date">

X00 means 100, 600, 700 and 800, X10 means 110, 610, 710 and 810.

Availability of Chronological Metadata Records

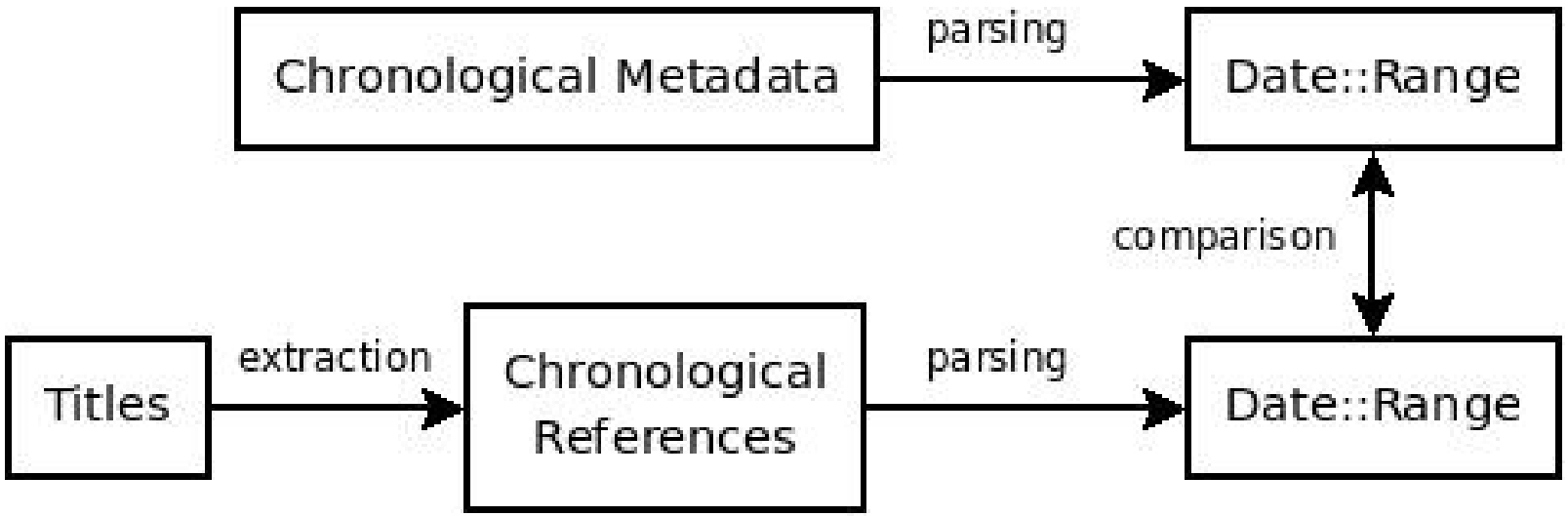
For each MARC file, we calculated the frequency of occurrence of each chronological MARC field.

	cwar	fsabw	gmd	mesnbib	papr
260$c	0.996	1	0.961	1	1
260$g			0.002		
046					
100$d	0.334	0.352	0.252		0.068
600$d	0.040	0.000	0.003		0.149
700$d	0.010	0.001	0.103		0.378
110$d			0.000		
610$d					0.004
710$d			0.000		0.001

We chose MARC 260$c (date of publication) as our work focus.

Outline

Overview of Chronological Metadata Remediation



Format Consistency

- Fixed length date format doesn't have consistency issue since the function of every character position is defined by the MARC standards.

- The format consistency of textual dates is not good. The MARC guidelines don't say much about the textual date format:

> Multiple adjacent publication dates such as a date of publication and copyright date are recorded in a single subfield $c. In records formulated according to ISBD principles, subfield $c is always preceded by a comma (,) unless it is the first subfield in field 260. Subfield $c ends with a period (.), hyphen (-) for open-ended dates, a closing bracket (]) or closing parenthesis ()). If subfield $c is followed by some other subfield, the period is omitted.

Parsing textual dates is the core part of chronological metadata remediation.

A List of Date Formats

We found various date formats in MARC 260$c from the metadata files we received. Here is an highly incomplete list of them:

Format	Instance
###-	[180-]
####-##	1601-15.
####-####	1862-1863
anno ####	anno 1668.
an V ####	an V. (1797)
between #### and ####	[between 1755 and 1762]
Bunka # ie ####	Bunka 1 i.e. 1804]
Decr # ####	Decr. 1, 1783.
#### SEASON?	1939 Spring?
Guangxu ## ####	Guangxu 30 [1904]
United States	[United States,

Parsing Textual Dates

Due to the variety of date formats, the parser for dates and date ranges should be very flexible and allow quick adding, modifying and removing parsing rules.

Perl modules:

- Date::Record. A single date object with attributes adopted from MODS.
- Date::Range. A date range object, equivalent to a MODS date element.
- Date::Format. A list of parsing rules with priority levels and handlers that output Date::Range objects.

Remediation Results

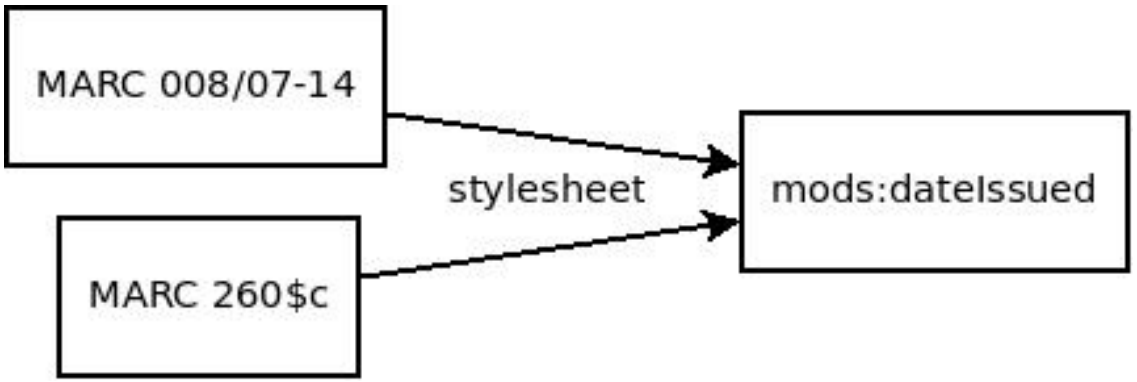
Abraham Lincoln Papers:

- 20158 records in non-MARC format.
- All the records contain <date_sorter>.

We parsed the titles and found 20127 titles containing dates and compared those dates with the corresponding <date_sorter> fields. There were only 13 mismatches, most of which were typos, e.g.:

date_sorter	item_title
1961-04-01	Abraham Lincoln, [April 1, 1861]
1863-01-14	Henry Grider to Abraham Lincoln, January 14, 1862
18650035	William B. Parkinson to Abraham Lincoln, March 5, 1865
185000930	Charles F. Anderson, Monday, September 30, 1850

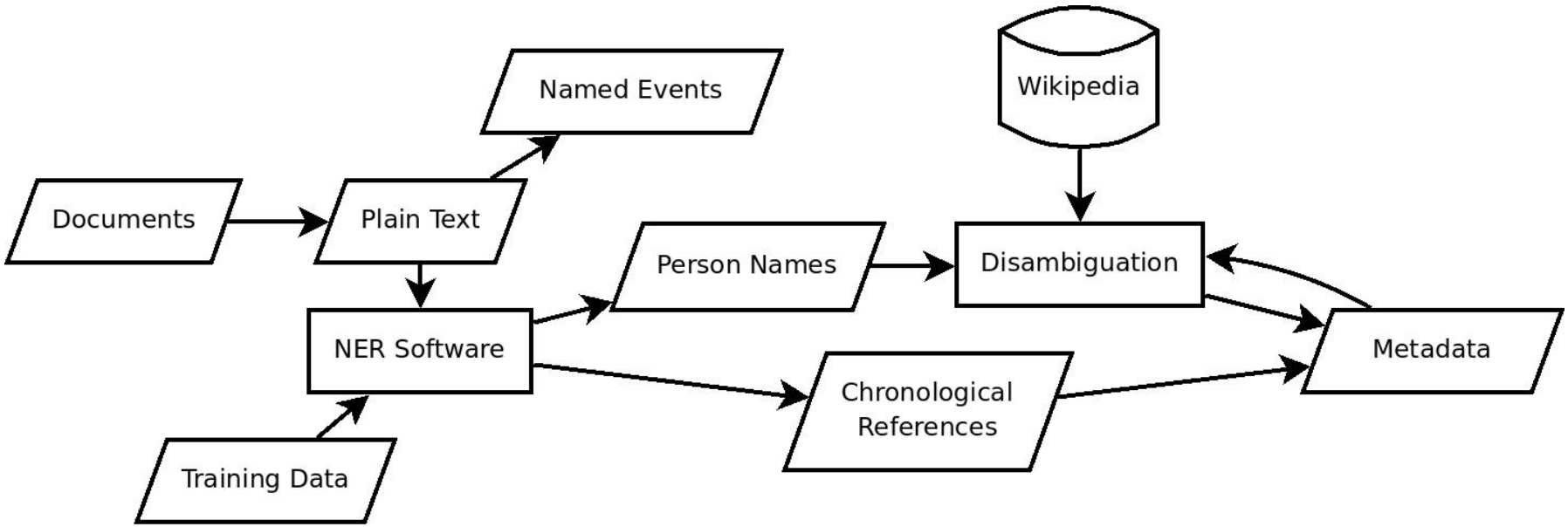
Removing Duplicate Records



Conversion from MARC or non-MARC to MODS may create duplicate records. For example, MARC 008/07-14 and 260$c usually contain the same date range. However, the conversion stylesheet won't check this and thus produce duplicate records. We suggest a post-processing step to remove the duplicates.

Outline

Overview of Chronological Metadata Augmentation



- NER software: Stanford Named Entity Recognizer.
- Training data: Annotated news wire articles from MUC-7.
- We use Freebase as a proxy to Wikipedia.

Future Works

- Write more parsing rules.
- Create a standard format for recording dates and date ranges?
- Deliver our codes:
  - Portable
  - User friendly (GUI and documentation)
- Utilize named events and person names for chronological metadata remediation.
- Build a prototype website.