The Evolution of MetaMap, A Concept Search Program for Biomedical Text

Alan (Lan) R. Aronson

François M. Lang

AMIA 2009

S40: Semantic Modeling and Mapping

November 16, 2009

Outline

- Background
  - Mapping programs
  - MetaMap distribution modes
  - Applications using MetaMap
- Recent MetaMap development
  - Tokenization issues
  - Output formats
  - Genre and task issues
  - Algorithm tuning

Historical Background

- Programs that map biomedical text to a thesaurus
  - CLARIT (*Evans et al., 1991*)
  - SAPHIRE (*Hersh et al., 1990*)
  - **MetaMap *(Aronson et al., 1994)***
  - Metaphrase (*Tuttle et al., 1998*)
  - **MMTx (*2001*)**
  - KnowledgeMap (*Denny et al., 2003*)
  - Mgrep (*2009*)
- Characteristics of MetaMap/MMTx
  - Linguistic rigor
  - Flexible partial matching
  - Emphasis on thoroughness rather than speed

MetaMap Example

- PMID – 19529903

- TI –Bile duct stricture due to caused by portal

  Stricture of bile duct | Causing | Hepatic

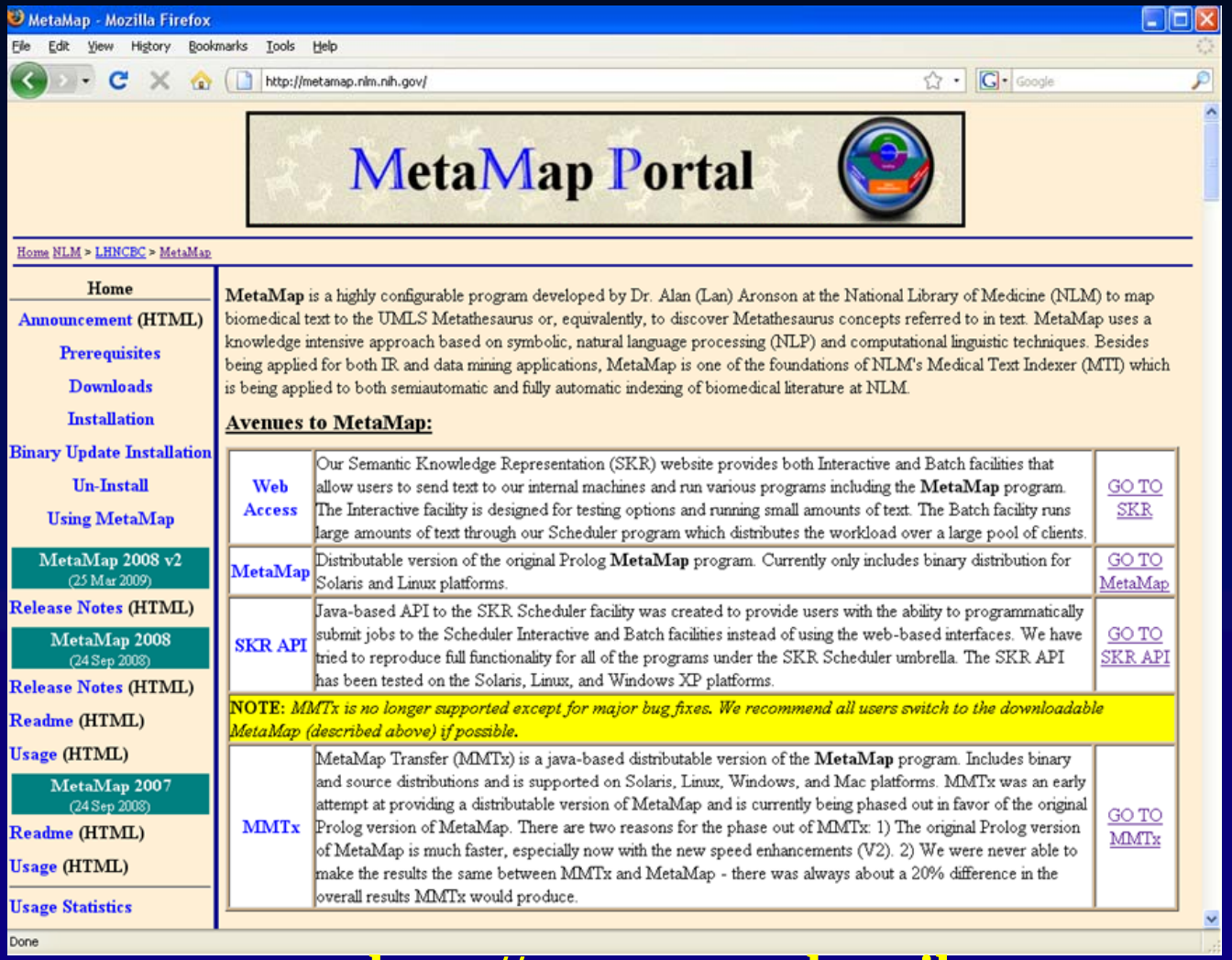
  biliopathy: Treatment with one-stage

  Administration procedure | One | Phase

  portal-systemic shunt and biliary bypass.

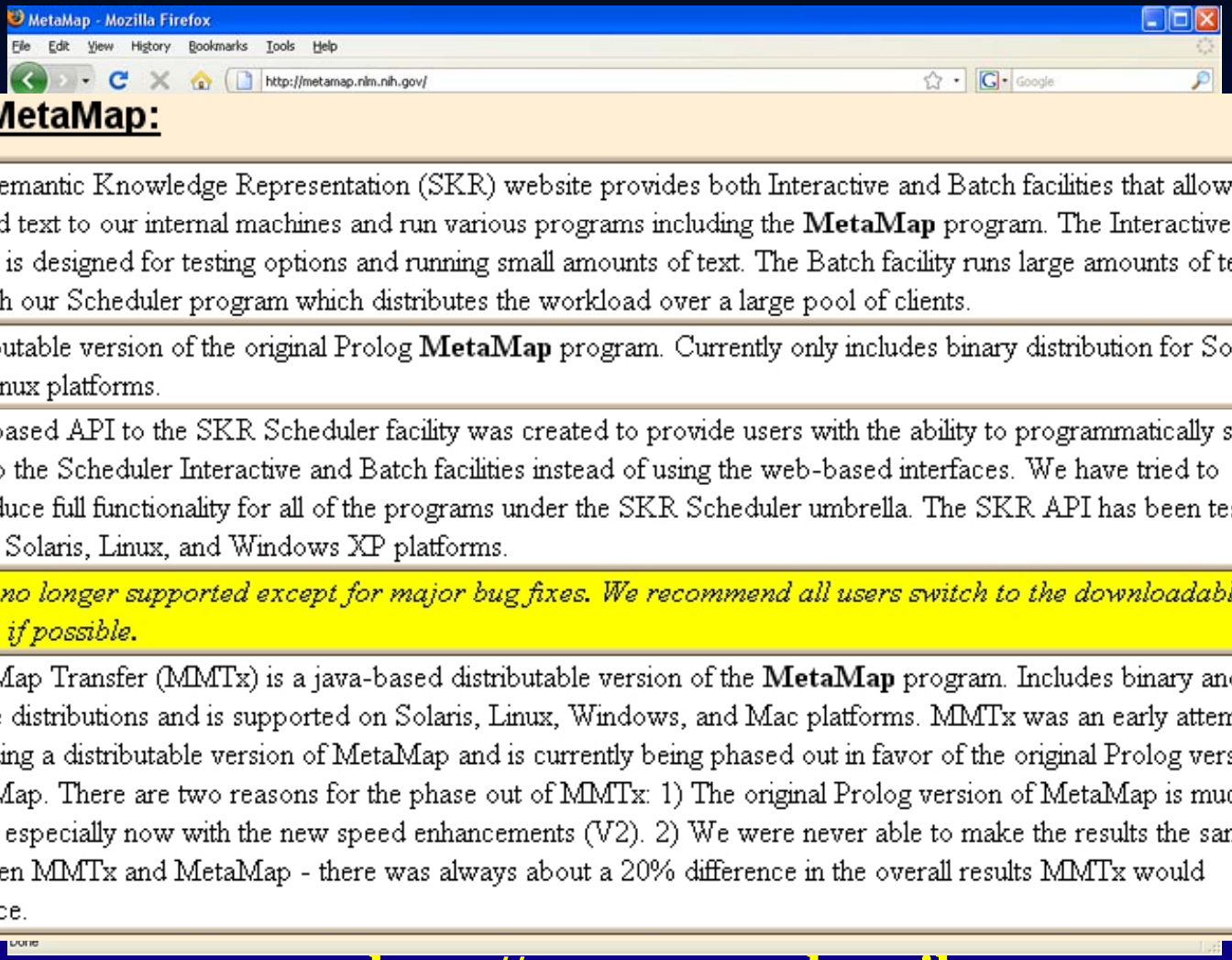
  Portasystemic shunt | Biliary | Bypass

MetaMap/MMTx Distribution Modes



http://metamap.nlm.nih.gov

MetaMap/MMTx Distribution Modes



Avenues to MetaMap:

Web Access	Our Semantic Knowledge Representation (SKR) website provides both Interactive and Batch facilities that allow users to send text to our internal machines and run various programs including the **MetaMap** program. The Interactive facility is designed for testing options and running small amounts of text. The Batch facility runs large amounts of text through our Scheduler program which distributes the workload over a large pool of clients.	GO TO SKR
MetaMap	Distributable version of the original Prolog **MetaMap** program. Currently only includes binary distribution for Solaris and Linux platforms.	GO TO MetaMap
SKR API	Java-based API to the SKR Scheduler facility was created to provide users with the ability to programmatically submit jobs to the Scheduler Interactive and Batch facilities instead of using the web-based interfaces. We have tried to reproduce full functionality for all of the programs under the SKR Scheduler umbrella. The SKR API has been tested on the Solaris, Linux, and Windows XP platforms.	GO TO SKR API
NOTE: *MMTx is no longer supported except for major bug fixes. We recommend all users switch to the downloadable MetaMap (described above) if possible.*		
MMTx	MetaMap Transfer (MMTx) is a java-based distributable version of the **MetaMap** program. Includes binary and source distributions and is supported on Solaris, Linux, Windows, and Mac platforms. MMTx was an early attempt at providing a distributable version of MetaMap and is currently being phased out in favor of the original Prolog version of MetaMap. There are two reasons for the phase out of MMTx: 1) The original Prolog version of MetaMap is much faster, especially now with the new speed enhancements (V2). 2) We were never able to make the results the same between MMTx and MetaMap - there was always about a 20% difference in the overall results MMTx would produce.	GO TO MMTx

http://metamap.nlm.nih.gov

NLM Applications using MetaMap

- Information retrieval (IR)
  - Indexing and query expansion experiments (*Aronson et al., Rindflesch et al.*)
  - Hierarchical indexing (*Wright, Grosetta-Nardini, et al.*)
  - TREC genomics track (*Aronson et al., Demner-Fushman et al., …*)
- Data mining
  - DAD (Drug-Adverse drug reactions-Disease) literature-based discovery (*Weeber et al.*)
  - Clinical findings (*Sneiderman et al.*)
  - Arbiter, EDGAR, anatomical terminology, SemRep, SemGen (*Rindflesch et al.*)
- NLM Indexing Initiative (II)
  - Medical Text Indexer (MTI) (*Aronson et al.*)
  - MeSH indexing experiment (*Kim, Aronson and Wilbur*)

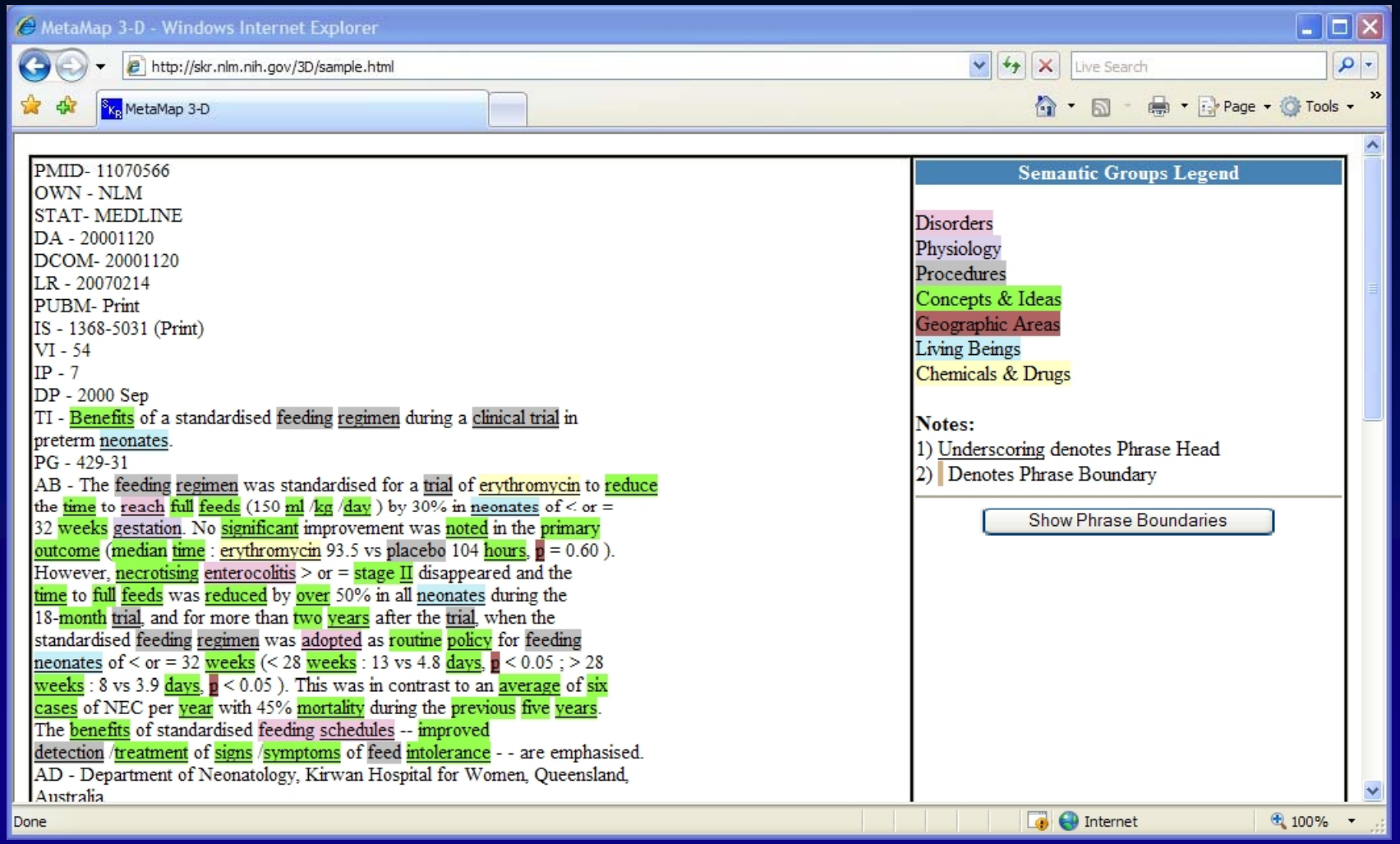
Tokenization Issues

- Acronym/abbreviation detection
  - e.g., “The effect of adrenocorticotropic hormone (ACTH) and cortisone on drug hypersensitivity reactions.”
  - Similar to Schwartz and Hearst, 2003 with rules:
    - AAs cannot contain > 20 characters
    - Single-word AAs cannot contain > 12 characters
    - …
- Non-standard input
  - e.g., several PubMed citations having no whitespace between sentences

Output Formats

- MetaMap Machine Output (MMO)
  - Prolog terms
  - Used for subsequent processing
- XML output
- Colorized MetaMap output (MetaMap 3D)

MetaMap 3D


MetaMap 3-D - Windows Internet Explorer
http://skr.nlm.nih.gov/3D/sample.html
Live Search
MetaMap 3-D
Page
Tools
PMID- 11070566
OWN - NLM
STAT- MEDLINE
DA - 20001120
DCOM- 20001120
LR - 20070214
PUBM- Print
IS - 1368-5031 (Print)
VI - 54
IP - 7
DP - 2000 Sep
TI - Benefits of a standardised feeding regimen during a clinical trial in preterm neonates.
PG - 429-31
AB - The feeding regimen was standardised for a trial of erythromycin to reduce the time to reach full feeds (150 ml /kg /day) by 30% in neonates of < or = 32 weeks gestation. No significant improvement was noted in the primary outcome (median time : erythromycin 93.5 vs placebo 104 hours, p = 0.60). However, necrotising enterocolitis > or = stage II disappeared and the time to full feeds was reduced by over 50% in all neonates during the 18-month trial, and for more than two years after the trial, when the standardised feeding regimen was adopted as routine policy for feeding neonates of < or = 32 weeks (< 28 weeks : 13 vs 4.8 days, p < 0.05 ; > 28 weeks : 8 vs 3.9 days, p < 0.05). This was in contrast to an average of six cases of NEC per year with 45% mortality during the previous five years. The benefits of standardised feeding schedules -- improved detection /treatment of signs /symptoms of feed intolerance - - are emphasised.
AD - Department of Neonatology, Kirwan Hospital for Women, Queensland, Australia
Semantic Groups Legend
Disorders
Physiology
Procedures
Concepts & Ideas
Geographic Areas
Living Beings
Chemicals & Drugs
Notes:
1) Underscoring denotes Phrase Head
2) | Denotes Phrase Boundary
Show Phrase Boundaries
Done
Internet
100%

Genre and Task Issues (1 of 2)

- Term processing (-z)
  - Input is terms (one per line), not complete sentences
- Browse mode (-zogm)
  - Used with Large Scale Vocabulary Text (LSVT)
  - Exhaustive search of the Metathesaurus
  - Voluminous output
  - Not appropriate for use with final mapping construction

Genre and Task Issues (2 of 2)

- Negation (--negex)
  - Important for clinical text
  - Based on Wendy Chapman's NegEx algorithm
- Word Sense Disambiguation (-y)
  - Based on Susanne Humphrey's Journal Descriptor Indexing
  - Provides modest improvement in results

Algorithm Tuning

- Variant suppression
  - Suppress variants of one- and two-character words
  - e.g., in *t-cell* suppressing variants of *t* prevents mapping to ‘TX’ and ‘TS’
- Efficiency modifications
  - Due to growth of Metathesaurus (440K – 2M concepts)
  - Caching results in AVL trees (self-balancing binary trees) rather than linear lists
  - Expanding caching scope from a phrase to a citation
  - Replacing findall/3 calls with recursive code
  - Significantly faster then before (at least 3-5 times)

Future MetaMap Development

- Further technical development
  - Migration from Sun/Solaris to Linux environment
  - Update to current Berkeley DB to prepare for
  - Migration from Quintus to SICStus Prolog
- Augment tokenization with chemical name recognition
- Enhance MetaMap's WSD accuracy with additional WSD algorithms
- Further enhancement of processing short words, especially acronyms/abbreviations

Pointers

http://metamap.nlm.nih.gov

Alan (Lan) R. Aronson (alan@nlm.nih.gov)
François M. Lang (flang@mail.nih.gov)