Name disambiguation and entity resolution

Techniques and tools for name disambiguation and entity resolution

Tools already actively in use in the VIVO community

prior to ingest

- Google Refine, now Open Refine

during ingest

- VIVO Harvester name matching, Pubmed Example Script
- XSLT ingest example

after ingest in VIVO

- URI Tool for cleaning data after import

Additional tools that may prove useful

borrowed from http://rawpatentdata.blogspot.com/2013/01/datamining-and-entity-resolutions-some.html

Name Cleaver

Name Cleaver (http://sunlightlabs.com/blog/2011/name-standardization-name-cleaver/) supports three major name types, politicians, individuals and organizations, with a specific class and special features for each.
The OrganizationNameCleaver class has methods to reduce a name to only the "kernel" of the name, and also to expand all abbreviations (that Name Cleaver knows of), useful for matching tasks.
The pyton code of the program can be downloaded here: https://github.com/sunlightlabs/name-cleaver

OYSTER entity resolution

OYSTER (Open sYSTem Entity Resolution) is an entity resolution system that supports probabilistic direct matching, transitive linking, and asserted linking. To facilitate prospecting for match candidates (blocking), the system builds and maintains an in-memory index of attribute values to identities. Because OYSTER has an identity management system, it also supports persistent identity identifiers. OYSTER is unique among other ER systems in that it is built to incorporate Entity Identity Information Management (EIIM). OYSTER supports EIIM by providing methods that enforce identifiers to be unique among identities, maintain persistent IDs over the life of an identity, and allowing the ability to fix false-positive and false-negative resolutions, which cannot be done with matching rules, through the use of assertion, traceability, and other features.
Developed in JAVA, can be downloaded from: http://sourceforge.net/projects/oysterer/

OPENCALAIS

Calais Web Service by Thomson Reuters. The web service is an API that accepts unstructured text (like news articles, blog postings, etc.), processes them using natural language processing and machine learning algorithms, and returns RDF-formatted entities, facts and events.
OpenCalais supports three types of entity disambiguation: Company disambiguation, Geographical disambiguation and Product (Electronics) disambiguation.
Disambiguation of company names - such as determining whether the company Olympus refers to Olympus Optical Co. Ltd. or Olympus Life and Material Science Europa. The resolution output for a given company mention includes:

- A URI that is unique and uniform across documents
- The formal English legal name of the company
- The company's ticker symbol (for public companies)

For company names that cannot be disambiguated, the returned results will include no resolution information.

AgroTagger

http://aims.fao.org/agrotagger
Used for indexing information resources, Agrotagger is a keyword extractor that uses the AGROVOC thesaurus as its set of allowable keywords. It can extract from Microsoft Office documents, PDF files and web pages.
Agrotagger began as a collaboration with Indian Institute of Technology of Kanpur (IITK) in 2010. Building on top of the popular Keyword Extraction Engine (KEA) the team created several versions, some based on a reduced subset of AGROVOC known as AGROTAGS (produced by partner ICRISAT) and others using the full set of AGROVOC concepts.