

# Information infrastructure for text mining of the primary scientific literature: Principles and Practice.

Gully Burns, Donghui Feng, Tommy Ingulfsen, Eduard Hovy  
Information Sciences Institute, University of Southern California  
{burns, donghui, tommying, hovy}@isi.edu

## Abstract

*We present an informatics infrastructure for biocuration, based on a combination of techniques from Information Extraction (IE) and Knowledge Engineering (KE). We describe the high-level design of this infrastructure which we base on the concept of ‘experimental type’. Here, we treat each experiment as a specific type of knowledge statement determined by the experiment’s design. We provide a preliminary, detailed example of the use of the infrastructure to support the construction of a database pertaining to neuroanatomical tract-tracing experiments. This work could generalize to provide support for other experimental types and could be used to make biocuration efforts more efficient. We also discuss how the process of annotating text for IE directly supports designing schema for databases.*

## 1. Introduction

The biological literature is probably the largest, most pervasive, most important, and yet most intractable collection of biomedical data currently available. Many large-scale databases (e.g., the ‘model organism’ databases listed at <http://gmod.org/>) are wholly based on information that has been curated from the literature. ‘Biocuration’ can be described as the task of constructing structured database entries from available accounts of experimental data, usually from the textual narrative of published papers. The efficient execution of this task is crucially central to the informatics infrastructure of modern biology [1], and developing automated tools to assist with this process forms the basis of the work described here.

**The literature as ‘semi-structured data’:** There are two main types of published scientific paper: primary experimental reports and reviews. The structure of experimental reports is quite regular and typically has the following sections: abstract, introduction, materials and methods, results, discussion / conclusion, and references. In comparison, the structure of review articles is freeform and is based mainly on citations linking to knowledge found in experimental reports (or other reviews). We focus on the originating source of new scientific knowledge by only considering primary research articles and disregarding reviews.

**‘Experimental Type’:** The literature itself is primarily a resource designed with human readability and retrieval in mind. Papers are not separated into information-specific categories that then may be collated into appropriate knowledge bases. To assist with this, we define the notion of ‘experimental-type’ as a guiding principle. Every scientific experiment has a structure that conforms to the principles of good experimental design (see [2] for a well-written guide to this process), but this structure is usually only reported implicitly within the description of the experimental methods and results. All of the seemingly complex choices made by an experimentalist to select a model system, methodology, assaying technique, time-points, and type of experimental subject are ‘independent variables’ and their values. All measurements made within the experiment are ‘dependent variables’ and their values. The ‘experimental type’ can therefore be defined by a particular configuration of independent and dependent variables.

Importantly, these same variables determine the schema used to represent the data provided by such an experiment. We might also ignore some of the less significant choices made by the experimenters and to use the ‘minimum information required by an experiment’. This particular idea has formed the basis of standardized object models for specific experiment types to enable collaboration and data sharing [3-6]. This natural parallel between the designs of experiments and schemas for bioinformatics systems prompts and enables our work.

**Information Extraction Infrastructure:** Given the size, complexity, and importance of the literature in biomedical research, Information Extraction (IE) is an important emerging research topic. Most work focuses on specific challenge problems such as the extraction of named biological entities or relations from text [7]. Here, we examine the overall problem of extracting composite entities, attributes, and relations from text to construct complete database tuples in a supervised learning framework. By necessity, this work is based on several components that must be brought together. We present an IE infrastructure that provides a general-purpose framework for the construction and population of subject-specific biomedical databases. In this paper, we discuss the principles underlying our strategy combined with an example for a specific type of biological experiment:

neuroanatomical tract-tracing. We anticipate that this framework could be used to construct and populate biomedical databases for other experimental types.

Figure 1 shows a simple view of the infrastructure’s design. As shown, our approach is based on standard supervised-machine-learning IE, relying on three actions to drive the generation of structured mark-up within text (which then provides the input for biomedical databases, see later): (a) the definition of a suitable schema, (b) annotation by domain experts, (c) automated markup of a large text corpus by a trained machine-learning system. The flowchart shows how evaluating the performance of this supervised learning process drives the refinement of the schema used to represent the experimental design, or add to the size of the human-annotated data used to train the system and scale up performance.

In this paper, we present how the use of this relatively simple infrastructure can drive the development of structured databases to reflect the semantic structure of textual data.

## 2. KE within IE: the ‘Analysis-Annotation’ Cycle

IE requires a suitable target representation or *schema* for the extraction. This often requires an *annotation manual* to assist workers performing the annotation. Naturally, the development of the schema and manual is iterative, as errors are encountered and must be corrected within the IE machinery. Furthermore novel linguistic constructs require new guidelines to capture the meaning when encountered in the text. During this iterative process, the design of the target representation may be adjusted according to Knowledge Engineering (KE) principles. An example serves to illustrate this point (see Figure 2).

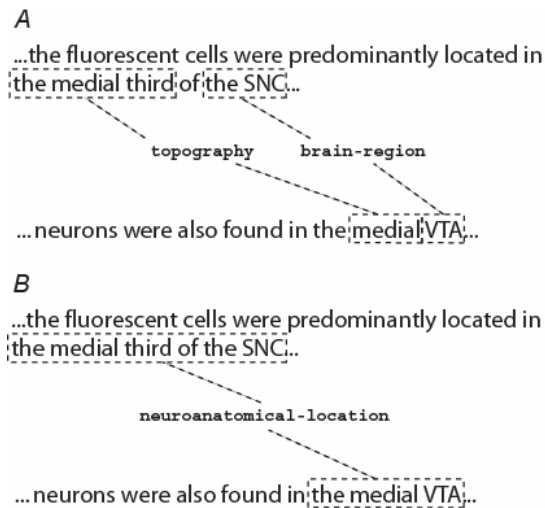


Figure 2: Schema refactoring during IE.

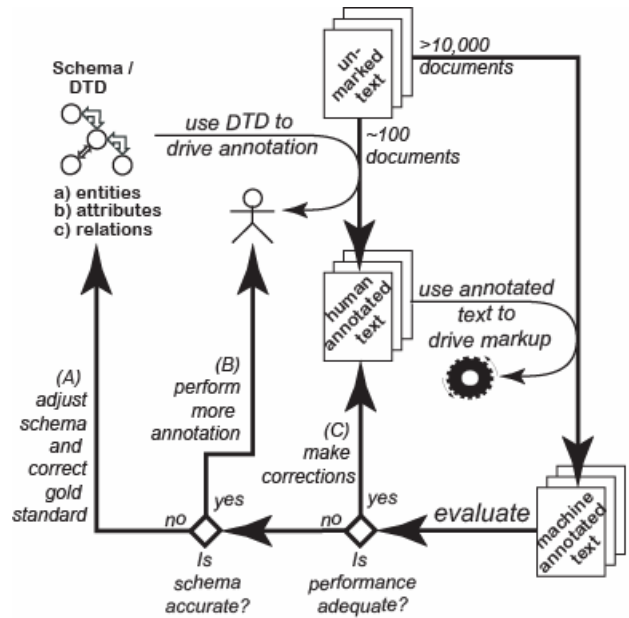


Figure 1: Simple IE infrastructure based on supervised learning.

This example consists of two sentences taken from [8]. Case A reflects our original markup design where we specified two fields. The first was **brain-region** with the text showing the SNC (Substantia Nigra, pars Compacta) and the VTA (Ventral Tegmental Area). The second was **topography** reflecting a smaller portion of tissue in one sector of the region. From the perspective of the target representation, each field denoted a separate attribute entry, but we felt the performance of the machine learning system under this markup scheme could be improved (F-Score ~ 0.75). Following a detailed re-evaluation of the text and our markup scheme, we redesigned the annotation to Case B, amalgamating the two attributes (**topography** and **brain-region**) to define a composite entity (**neuroanatomical-location**). After performing this step, performance appeared to improve (F-Score ~ 0.79), due, in part, to the presence of very common textual features such as the word ‘in’ which preceded each instance of **neuroanatomical-location**.

Although this type of refinement is probably very commonplace for supervised IE work in general this small, seemingly irrelevant step is significant in the context of biomedical database development. Annotating the text of the literature for IE provides valuable feedback for database design. This feedback emerges from common usage of concepts by authors within the primary literature. Designing and redesigning markup schema for the IE that accommodates the complexity of these concepts is *exactly* what is required by knowledge engineers designing biomedical databases. The IE infrastructure can also accelerate the population of databases from the literature.

### 3. A Worked Example: IE for Tract-Tracing Experiments

#### 3.1. Input Data

**Acquisition:** IE approaches usually work as supervised or semi-supervised machine-learning methods where raw text must first be annotated to provide learnable features. We acted within publishers' copyright restrictions to create a text corpus containing roughly 12,000 articles, dating from 1970 to 2005 of the Journal of Comparative Neurology (JCN, an authoritative neuroanatomical journal [9]). We extracted (x,y) coordinates and formatting of each word to provide additional cues to perform coarse-grained decomposition to classify the documents' sections ('Introduction', 'Results', etc.). We performed IE only on Results sections.

**Schema design:** We drew on previous knowledge engineering work in neuroinformatics [10, 11] to devise a preliminary design for our target schema for IE. An individual tract-tracing experiment only requires four separate pieces of information to define an interpretable representation of the experimental data. These are (A) **injection-location** (the location where tracer chemical was originally deposited), (B) **labeling-location** (the location where tracer was transported to along axonal fibers), (C) **tracer-chemical** (the type of tracer which denotes the direction of transport), and (D) **labeling-description** (the type and density of labeling, denoting the strength and nature of the connection) [12].

**Annotation:** XML is widely used to support language processing annotation. We use Vex, a free XML editor [13] based on the Eclipse platform, to perform the annotation task. Vex provides configurable control over the appearance of the text with different colors for different tags and classes. Vex permits annotators to make corrections and to select tags via simple right-click functionality, speeding up annotation significantly.

#### 3.2. Processing

**Conditional Random Fields Labeling:** The Conditional Random Field (CRF) model for sequential labeling [14] has been widely used in many aspects of language processing (improved model variants, [15], web data extraction [16], scientific citation extraction [17], and word alignment [18]). The model's originators provide an open-source toolkit (MALLET [19]) which forms the basis of our approach. CRF models are very versatile [20] and can accommodate a wide range of different types of textual features.

**Feature Functions:** The CRF model relies on sets of features to provide the cues that it bases its classification on. The choice of a good set of feature is crucial for IE systems to function well. In this case, we use four types

of feature function: (i) Lexical knowledge based on domain-specific terms, (ii) unigrams based on the current, preceding or following word, (iii) syntactic dependency, (iv) user-defined specific semantic context (in this study: "does the word 'injection' appear in this sentence?", we call these 'context window' features).

**Active Learning:** One aspect of the process of developing workable IE infrastructure is to minimize the amount of text annotation required, as much as possible. 'Active learning' approaches seek to prioritize the annotation task so that the most beneficial texts are annotated first [21]. Thus, in an incremental paradigm such as ours, where we train our model, run it on unmarked data, and then make corrections before adding marked up texts to our training data, this methodology may improve performance of the overall infrastructure.

We integrated an uncertainty-based active learning framework with the CRF model to implement this approach. We used two heuristic methods (based on 'peer' and 'set' comparisons) to determine certainty scores to estimate the quality of given labeling sequence's quality. Peer comparisons are made between the highest-scoring and highest-scoring-but-one labeling sequence. By contrast, set comparisons are made between the highest scoring labeling sequence and a set of N-best sequences.

**Evaluation:** We used Precision, Recall, and F-Score to evaluate the correctness of meaningful field values. These measures are defined below in Equations 1, 2, and 3.

$$Precision = \frac{\# \text{ of correct meaningful labels}}{\# \text{ of the meaningful labels by the system}} \quad (1)$$

$$Recall = \frac{\# \text{ of correct meaningful labels}}{\# \text{ of the meaningful labels in the gold standard}} \quad (2)$$

$$F - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

#### 3.3. Results

**Study design:** To study the basic performance of our IE informatics infrastructure we selected and hand-annotated 21 research articles from the JCN corpus with tract-tracing XML tags. These articles' results sections contained a total of 2009 individual sentences, of which only 1029 sentences were annotated with one or more of the tags described in section 3.1. We split the text into training and testing data according to a 2:1 ratio (685 training and 344 testing sentences).

**IE Performance:** To establish a baseline system, we scanned every sentence for words or phrases from each lexicon. If the term was present, then we labeled the word based on the lexicon in which it appeared. If words appeared in multiple lexicons, we assigned labels randomly. We tested performance under different

combinations of features. All feature combinations performed higher than baseline (see Table 1)

**Table 1: NLP performance (Precision, Recall and F-Score) for text-mining from tract-tracing experiments. Features Key, ‘L’ = Lexical, ‘C’ = Current Word, ‘P/F’ = Preceding or Following Word, ‘W’ = ‘Context Window’, ‘D’ = Dependency features.**

Features	Precision	Recall	F-Score
Base	0.41	0.18	0.25
L	0.60	0.37	0.46
L + C	0.77	0.73	0.75
L + C + W	0.77	0.73	0.75
L + C + W + P/F	0.81	0.75	0.78
L + C + W + P/F + D	0.80	0.78	0.79

Performance of this task is acceptably high (F-Score = 0.79). This is especially encouraging because the number of training examples (approximately the equivalent of 14 documents) is relatively small. We then ran our classifier on previously unseen text and corrected the markup. We found that annotation had accelerated from approximately 45 sentences per hour (sn/hr) to a rate

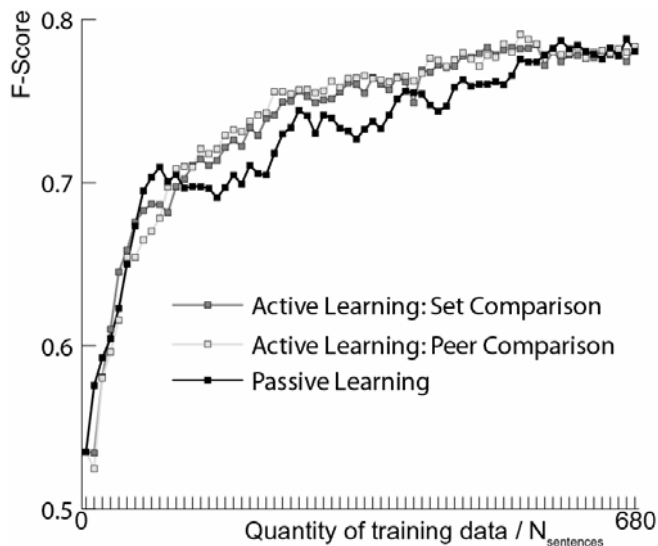
Counts	human labels							
	O	Location	Injection Spread	Injection Description	labeling Location	labeling Chemical	tracer	
O	41087	141	97	338	1751	6	43420	
InjectionLocation	545	744	48	6	820	1	2164	
InjectionSpread	126	43	147	11	155	0	482	
labelingDescription	1121	5	0	3773	82	47	5028	
labelingLocation	1988	224	110	27	9251	0	11600	
tracerChemical	108	1	12	0	0	623	744	
	44975	1158	414	4155	12059	677		

**Figure 3: Confusion matrix for tract-tracing IE.** of 115 sn/hr.

A consideration of the *confusion matrix* describing the errors made by our system is revealing. In Figure 3, the leading diagonal holds the counts of our system’s correct guesses for word-labels, and off-diagonal counts demonstrate errors. Note that the three labels for different types of neuroanatomical locations are frequently confused (**injection-location**, **tracer-chemical**, and **labeling-location**). Pooling these labels into a single category yields  $Recall = 0.81$ ,  $Precision = 0.85$ ,  $F-Score = 0.83$ .

**Active Learning Results:** We selected a set of ‘seed’ data from the training pool and then iterated through our active-learning paradigm. At every step we trained a new CRF model and labeled sentences in the rest of the training pool. Those with the lowest ranked scores for certainty were selected for inclusion in the training set. We set this the number of examples to be included with each iteration to 10 so that we added these sentences with the lowest certainty scores to the training set. We demonstrate the potential increase in learning efficiency

in Figure 4, as indicated by the difference between F-Scores for the active and passive learning cases.



**Figure 4: F-Score within Active Learning Paradigm.**

**Constructing a useable system:** Finally, we scaled up the text-mining process to accommodate all available papers from the corpus. We have mined all available text in our corpus for tract-tracing data and have provided access to the data and text via a protected web-interface (for use by people within our institution, for copyright reasons). This database was built from a test set of 8911 files and contains 751,915 records. There are 31,428 injection-location instances, 17,669 tracer-chemical instances, 385,138 labeling-location instances, and 317,680 labeling-description instances. The interface allows users to search for data under each of the entity tags described in section 3.1 and provides access to the marked up XML documents. This interface is for use by curators of the Brain Architecture Management System [22] as a search tool for the tract-tracing literature.

## 4. Conclusion

In this paper, we present a computational infrastructure for a reconfigurable text-mining system to construct a database of experimental data from the published literature. We seek to use well-established computational techniques such as supervised learning, and active learning. We strategically combine them to provide support for biocuration that fits into existing processes of designing and populating existing systems. Importantly, we seek to provide a framework that supports the development of biomedical databases in subjects that currently are not supported with informatics systems. We envisage that the development of low-cost IE systems such as this may permit the creation of small-scale databases that significantly assist scholarly work in

the biosciences (see [23, 24] for preliminary software constructs to support this vision).

Of particular interest of this work is the possibility of the development of ontology-engineering methods from using IE annotation within the biocuration process. Large-scale bioinformatics database hire full-time biocuration staff (often at the Ph.D. level of expertise). The Jackson laboratories have over twenty full-time curators for the Mouse Genome Database (Janan Eppig, personal communication). If this staff were provided with an effective IE system, the coverage, depth and sophistication of all aspects (schema, ontological vocabulary and data) of the system would improve dramatically.

## 5. References

1. Bourne, P.E. and J. McEntyre (2006), "Biocurators: contributors to the world of science". *PLoS Comput Biol*, **2**(10): p. e142
2. Ruxton, G.D. and N. Colegrave (2003), "Experimental design for the life sciences", Oxford: Oxford University Press.
3. Brooksbank, C. and J. Quackenbush (2006), "Data standards: a call to action". *Omics*, **10**(2): p. 94-9
4. Brazma, A., *et al.* (2001), "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data". *Nat Genet*, **29**(4): p. 365-71
5. Deutsch, E.W., *et al.* (2006), "Development of the Minimum Information Specification for In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)". *Omics*, **10**(2): p. 205-8
6. Leebens-Mack, J., *et al.* (2006), "Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA)". *Omics*, **10**(2): p. 231-7
7. Cohen, A.M. and W.R. Hersh (2005), "A survey of current work in biomedical text mining". *Brief Bioinform*, **6**(1): p. 57-71
8. Albanese, A. and D. Minciacchi (1983), "Organization of the ascending projections from the ventral tegmental area: a multiple fluorescent retrograde tracer study in the rat". *J Comp Neurol*, **216**(4): p. 406-20
9. Saper, C.B. (1999), "What's in a citation impact factor? A journal by any other measure". *J Comp Neurol*, **411**(1): p. 1-2
10. Burns, G.A. and W.C. Cheng (2006), "Tools for Knowledge Acquisition within the NeuroScholar system and their application to anatomical tract-tracing data". *J Biomed Discov Collab*, **1**(1): p. 10
11. Burns, G.A. (2001), "Knowledge management of the neuroscientific literature: the data model and underlying strategy of the NeuroScholar system". *Philos Trans R Soc Lond B Biol Sci*, **356**(1412): p. 1187-208
12. Blackstad, T., L. Heimer, and E. Mugaini (1981), "General approaches and laboratory procedures", in "Neuroanatomical tract tracing techniques", L. Heimer and M. Robads, Editors, Plenum Press: New York and London.
13. Vex - A Visual Editor for XML  
[\[http://vex.sourceforge.net/\]](http://vex.sourceforge.net/)
14. Lafferty, J., A. McCallum, and F. Pereira (2001), "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data". in *Proceedings of the International Conference on Machine Learning*.
15. Jiao, F., *et al.* (2006), "Semi-supervised conditional random fields for improved sequence segmentation and labeling." in *The annual meeting for the Association for Computational Linguistics (ACL-2006)*.
16. Pinto, D., *et al.* (2003), "Table Extraction Using Conditional Random Fields". in *Proceedings of the ACM SIGIR*.
17. Peng, F. and A. McCallum (2004), "Accurate information extraction from research papers using conditional random fields". in *Proceedings of HLT-NAACL*. p. 329-336.
18. Blunsom, P. and T. Cohn (2006), "Discriminative word alignment with conditional random fields." in *The annual meeting for the Association for Computational Linguistics (ACL-2006)*.
19. Mallet - Advanced Machine Learning for Language  
[\[http://mallet.cs.umass.edu/\]](http://mallet.cs.umass.edu/)
20. Sutton, C. and A. McCallum (2006), "An Introduction to Conditional Random Fields for Relational Learning." in "In Introduction to Statistical Relational Learning." L. Getoor and B. Taskar, Editors, MIT Press.
21. Thompson, C.A., M.E. Califf, and R.J. Mooney (1999), "Active learning for natural language parsing and information extraction." in *The Sixteenth International Conference on Machine Learning*. Bled, Slovenia.
22. Bota, M., H. Dong, and L.W. Swanson (2005), "The Brain Architecture Management System." *Neuroinformatics*, **3**(1): p. 15-48
23. Burns, G.A., *et al.* (2003), "Tools and approaches for the construction of knowledge models from the neuroscientific literature". *Neuroinformatics*, **1**(1): p. 81-109
24. Khan, A., *et al.* (2006), "NeuroScholar's Electronic Laboratory Notebook and its Application to Neuroendocrinology". *Neuroinformatics*, **4**(2): p. 139-160