CONTENT MINE

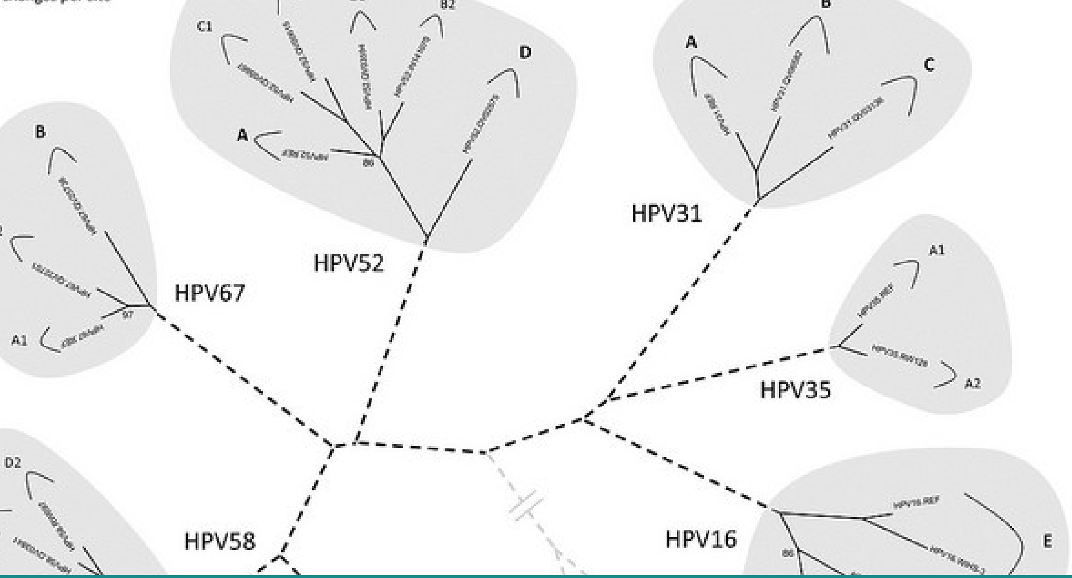
SHUTTLEWORTH FOUNDATION

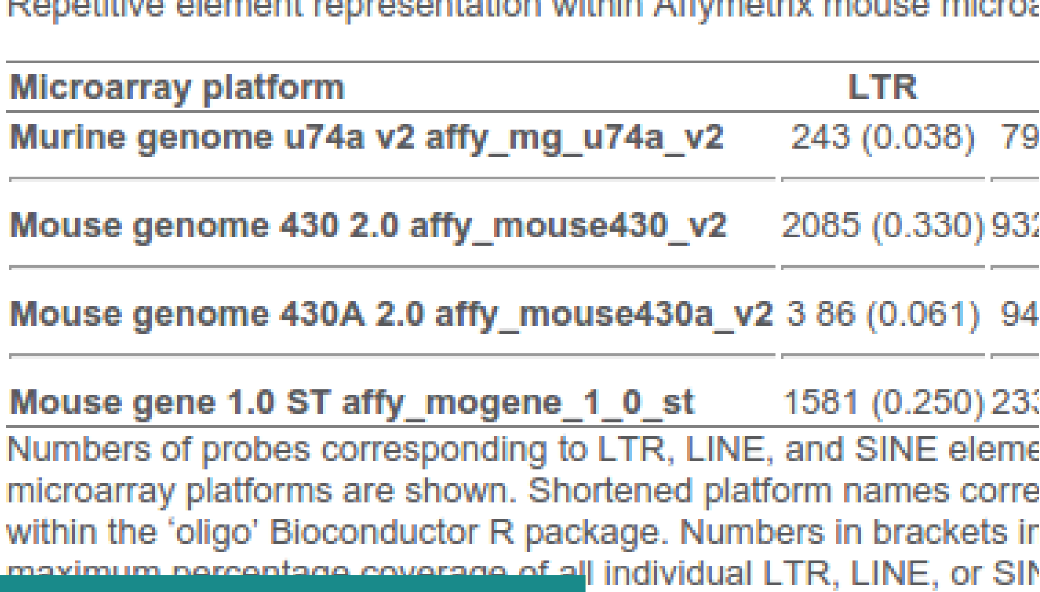
ContentMine:

extracting millions of facts from the scientific literature

@jenny_molloy

OpenCon Nov 2015

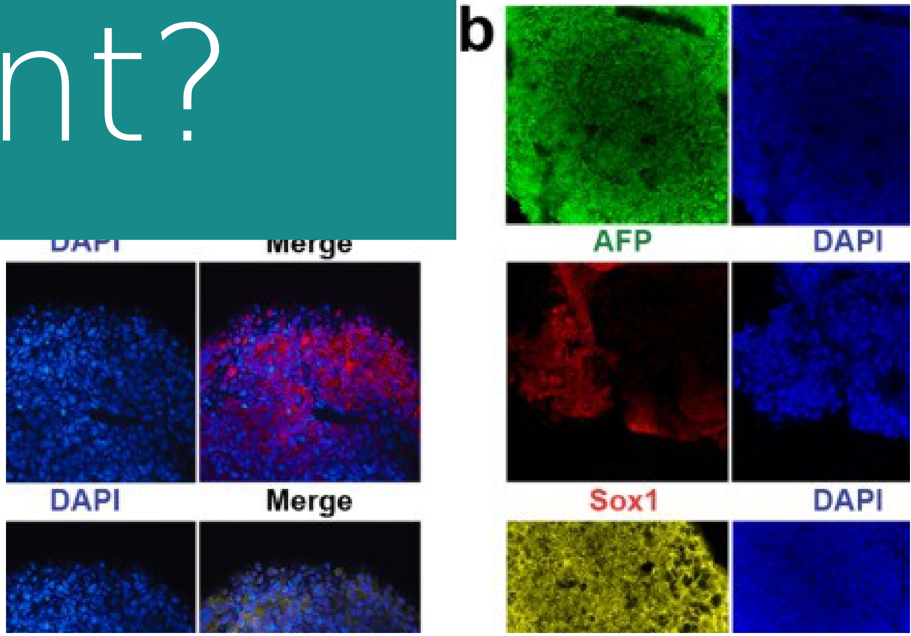


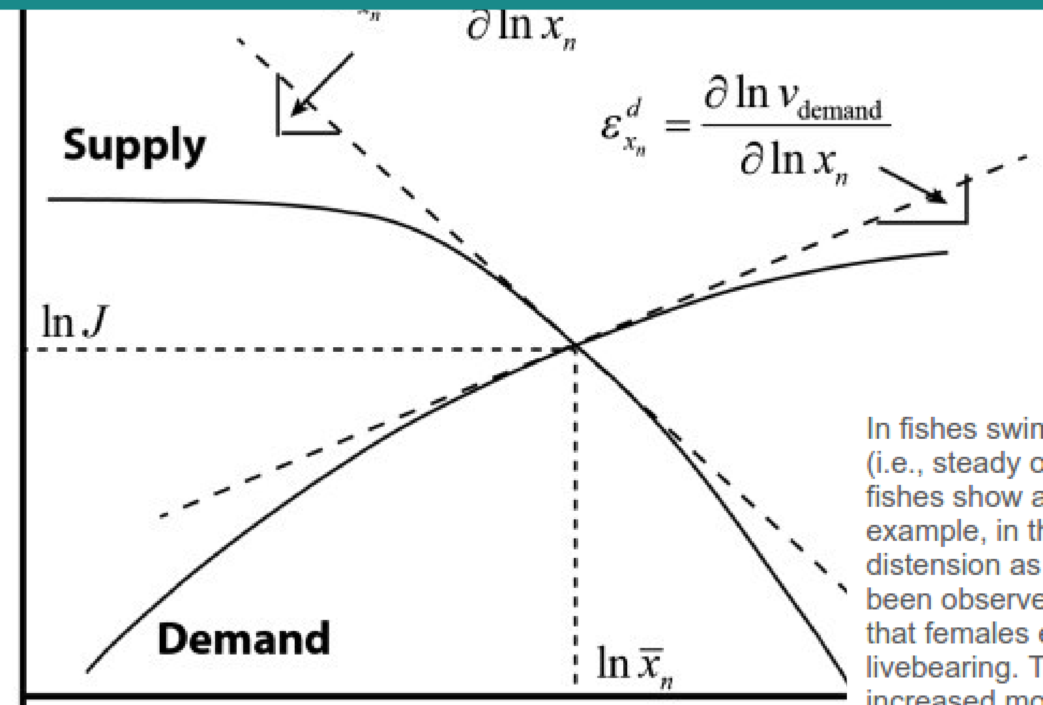

Repetitive element representation within Affymetrix mouse micro

Microarray platform	LTR	
Murine genome u74a v2 affy_mg_u74a_v2	243 (0.038)	79
Mouse genome 430 2.0 affy_mouse430_v2	2085 (0.330)	932
Mouse genome 430A 2.0 affy_mouse430a_v2	3 86 (0.061)	94
Mouse gene 1.0 ST affy_mogene_1_0_st	1581 (0.250)	233

Numbers of probes corresponding to LTR, LINE, and SINE eleme microarray platforms are shown. Shortened platform names corre within the 'oligo' Bioconductor R package. Numbers in brackets ir maximum percentage coverage of all individual LTR, LINE, or SIN

What is content?





In fishes swimming performance is a major determinant of survival probability [16]. Swimming a (i.e., steady or unsteady swimming measures) is strongly related to body form [16]–[19]. Livebe fishes show a change in body form and increased overall mass in the latter stages of pregnanc example, in the livebearing fish *Brachyrhaphis rhabdophora*, females exhibit increased abdomi distension as pregnancy progresses [20], [21]. Pregnancy-related reduction in escape velocity been observed in western mosquitofish (*Gambusia affinis*) [22], [23] and these studies have su that females experience a viability cost of reproduction as a consequence of the physical burde livebearing. The argument is that changes in shape lead to reduced swimming ability which lea increased mortality of pregnant females. However, the magnitude of this viability cost of reprod

1982

“**Automatically** generating logical representations of **text passages**… by means of an **analysis** of the coherence structure of the passages.”

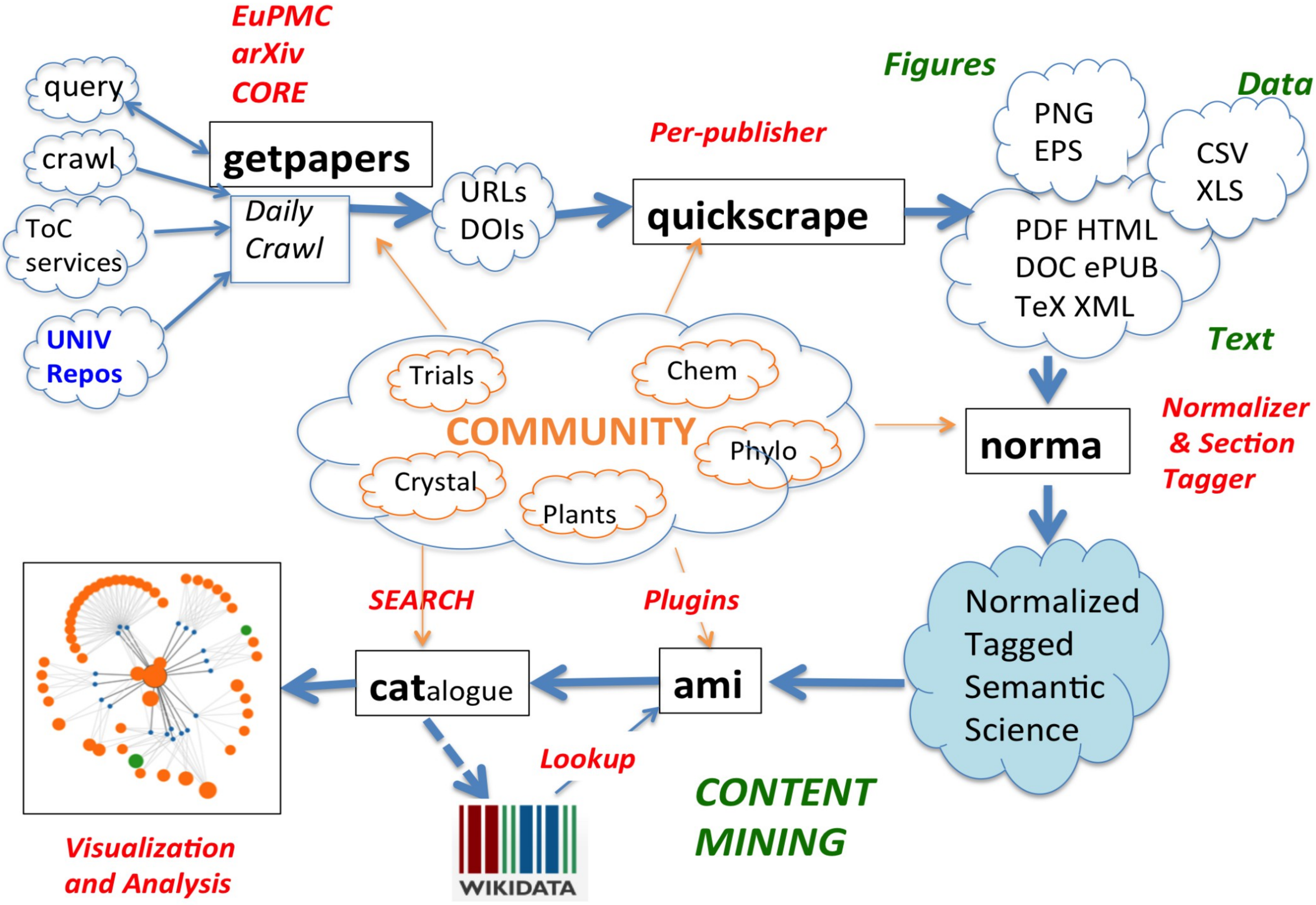
Jerry R. Hobbs, Donald E. Walker, and Robert A. Amsler. 1982. Natural language access to structured text. In Proceedings of the 9th conference on Computational linguistics - Volume 1(COLING '82), Ján Horecký (Ed.), Vol. 1. Academia Praha, , Czechoslovakia, 127-132. DOI=10.3115/991813.991833 http://dx.doi.org/10.3115/991813.991833

What is mining?

2008

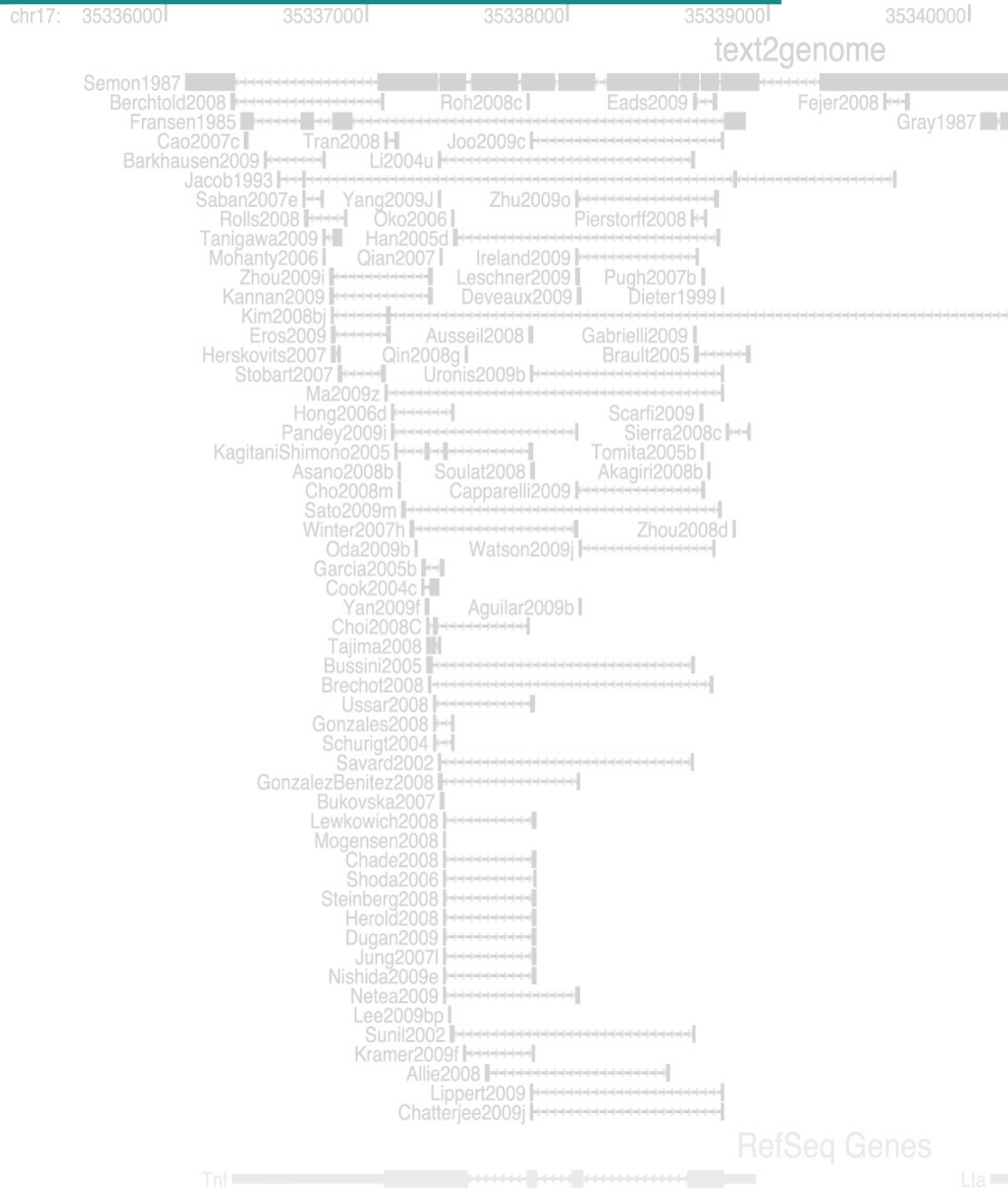
“The use of **automated methods** for exploiting the enormous amount of **knowledge** available in the biomedical literature.”

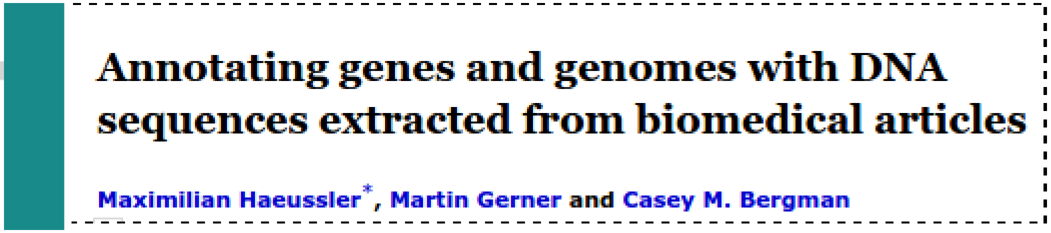
Cohen, K. Bretonnel; Hunter, Lawrence (2008). "Getting Started in Text Mining". PLoS Computational Biology 4 (1): e20. doi:10.1371/journal.pcbi.0040020. PMC 2217579.PMID 18225946.


EuPMC
arXiv
CORE
query
crawl
ToC services
UNIV Repos
getpapers
Daily Crawl
URLs DOIs
Per-publisher
quickscrape
Figures
PNG EPS
Data
CSV XLS
PDF HTML DOC ePUB TeX XML
Text
norma
Normalizer & Section Tagger
Trials
Chem
COMMUNITY
Phylo
Crystal
Plants
SEARCH
Plugins
Normalized Tagged Semantic Science
catalogue
ami
Lookup
WIKIDATA
CONTENT MINING
Visualization and Analysis

Annotation

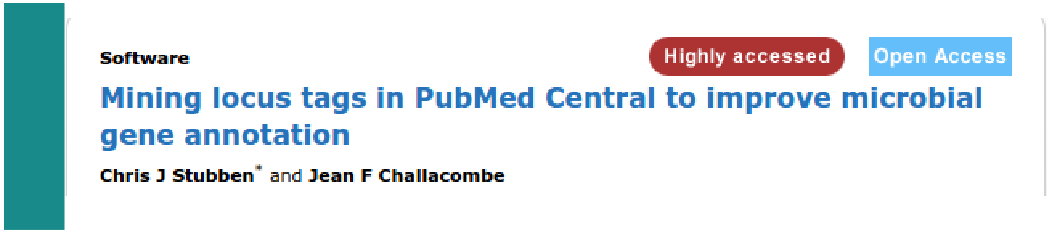
Potential to improve
quality and **efficiency**
of genomic research.





Annotating genes and genomes with DNA sequences extracted from biomedical articles

Maximilian Haeussler*, Martin Gerner and Casey M. Bergman



Software

Highly accessed

Open Access

Mining locus tags in PubMed Central to improve microbial gene annotation

Chris J Stubben* and Jean F Challacombe

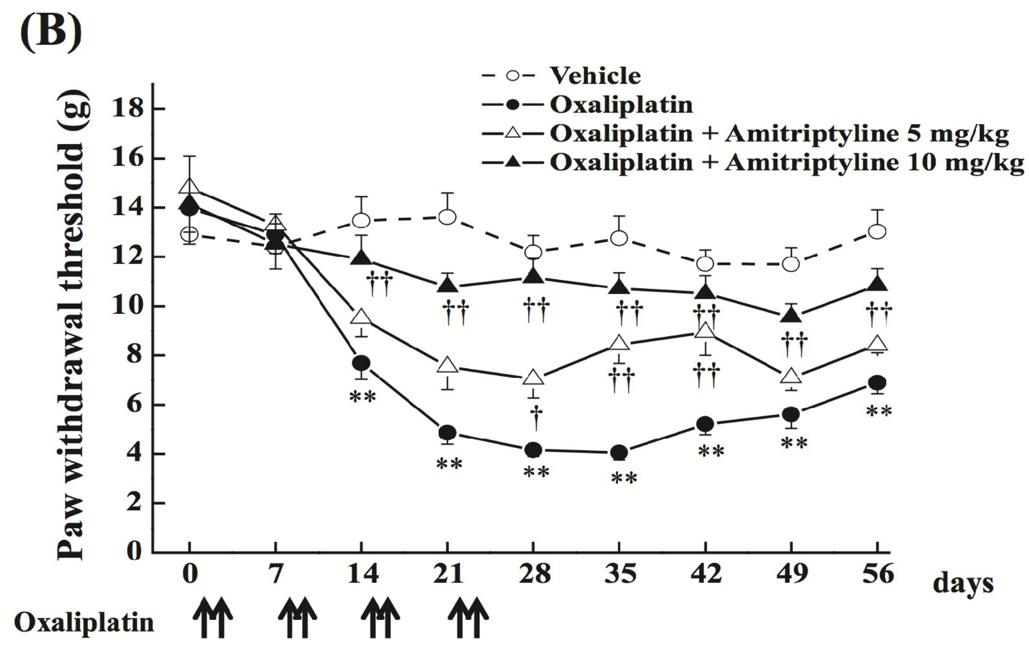
Many applications:

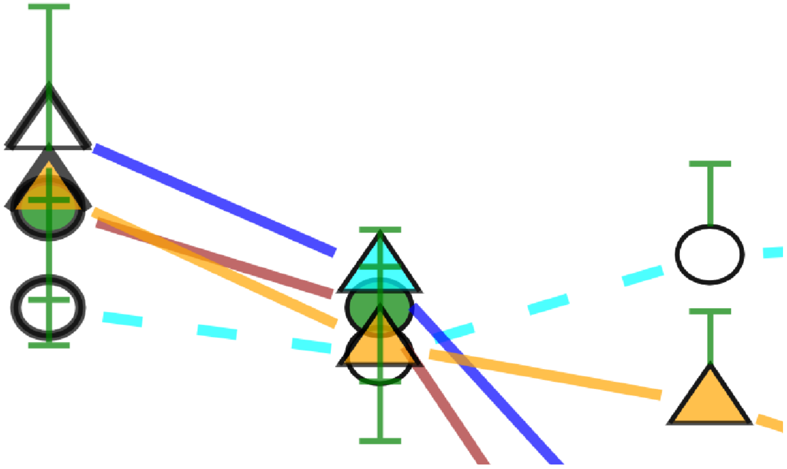
- Find primers
- Enhance positive controls
- Find novel sequence information
- More detailed and accurate annotation

Clinical Trials

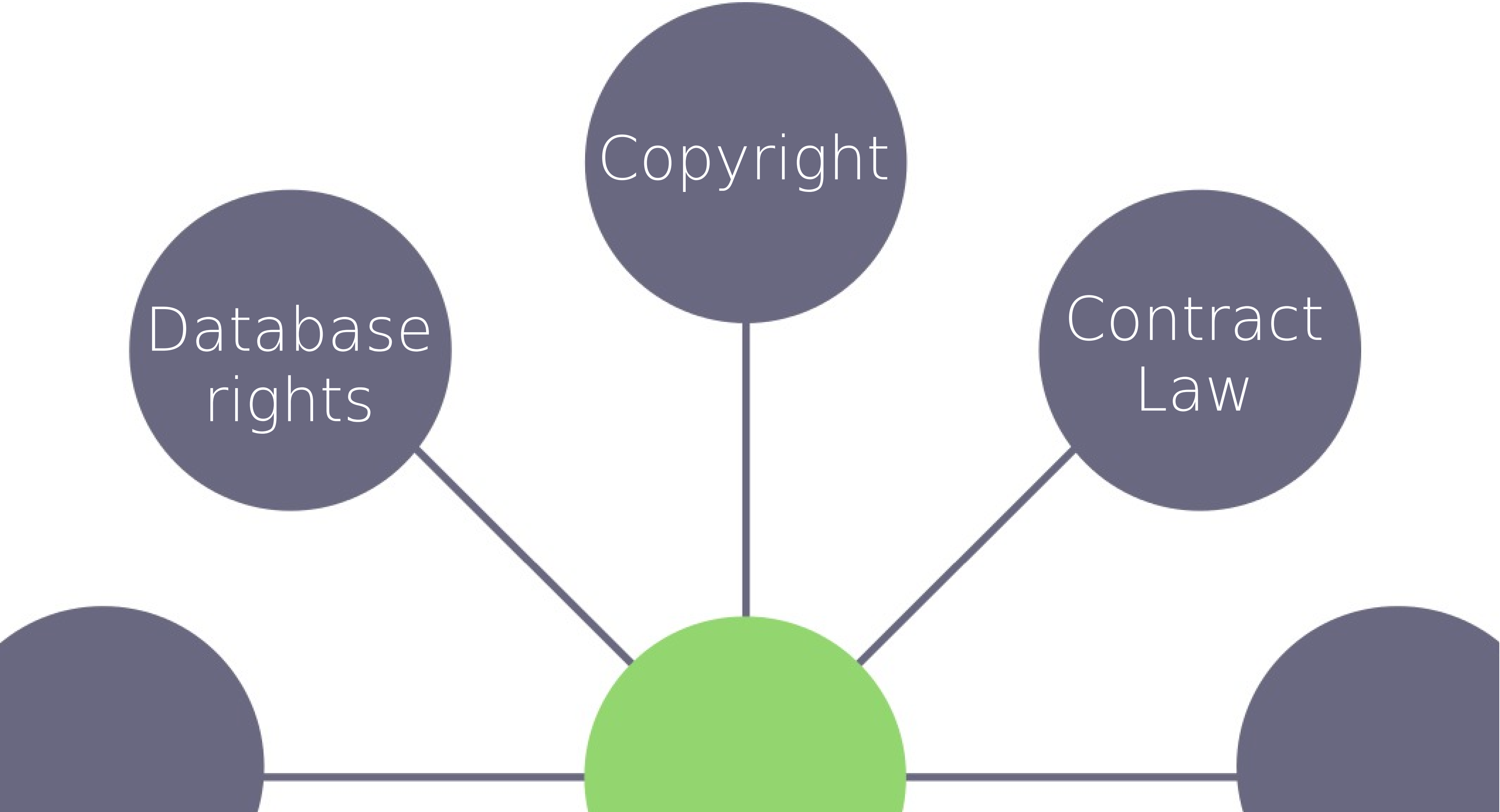
Clinical trials offer clear use cases for content mining.

Data extraction from graphs could be very useful for meta-analyses where raw data is unavailable.





Legal Considerations



bit.ly/cm-opencon15

Thank you very much
for your attention!
Any questions?

CONTENT
MINE

Peter Murray-Rust
Ross Mounce
Richard Smith-Unna
Jenny Molloy
Mark MacGillivray
Graham Steel
Stefan Kasberger
Christopher Kittel

With thanks to:
Charles Oppenheim
Michelle Brook

Follow
@TheContentMine

contentmine.org

Find the code on
github.com/ContentMine

Funded by:

SHUTTLEWORTH
FOUNDATION

What is Content?

Phylogenetic Tree from Figure 1 in Evolution and Taxonomic Classification of Human Papillomavirus 16 (HPV16)-Related Variant Genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. Chen Z, Schiffman M, Herrero R, DeSalle R, Anastos K, et al. (2011) Evolution and Taxonomic Classification of Human Papillomavirus 16 (HPV16)-Related Variant Genomes: HPV31, HPV33, HPV35, HPV52, HPV58 and HPV67. PLoS ONE 6(5): e20183. doi: 10.1371/journal.pone.0020183

Graph from He F, Fromion V, Westerhoff HV. (Im)Perfect robustness and adaptation of metabolic networks subject to metabolic and gene-expression regulation: marrying control engineering with metabolic control analysis. BMC Syst Biol. 2013;7 131. doi:10.1186/1752-0509-7-131. PubMed PMID: 24261908; PubMed Central PMCID: PMC4222491.

Table from Table 1 Young GR, Mavrommatis B, Kassiotis G. Microarray analysis reveals global modulation of endogenous retroelement transcription by microbes. Retrovirology. 2014;11 59. doi:10.1186/1742-4690-11-59. PubMed PMID: 25063042; PubMed Central PMCID: PMC4222864.

Text from Laidlaw CT, Condon JM, Belk MC. Viability Costs of Reproduction and Behavioral Compensation in Western Mosquitofish (Gambusia affinis). PLoS One. 2014;9(11) e110524. doi:10.1371/journal.pone.0110524. PubMed PMID: 25365426; PubMed Central PMCID: PMC4217728.

Cell microscopy image from Pettinato G, Vanden Berg-Foels WS, Zhang N, Wen X. ROCK Inhibitor Is Not Required for Embryoid Body Formation from Singularized Human Embryonic Stem Cells. PLoS One. 2014;9(11) e100742. doi:10.1371/journal.pone.0100742. PubMed PMID: 25365581; PubMed Central PMCID: PMC4217711.

Annotation:

Stubben, C. J., & Challacombe, J. F. (2014). Mining locus tags in PubMed Central to improve microbial gene annotation. BMC bioinformatics, 15(1), 43.

Figure from Haeussler, M., Gerner, M., & Bergman, C. M. (2011). Annotating genes and genomes with DNA sequences extracted from biomedical articles. Bioinformatics, 27(7), 980-986.