Deep Indexing and Discovery of Tables and Figures

Dr. Robert J. Sandusky
AUL for Information Technology
Richard J. Daley Library
University of Illinois at Chicago

NISO Discovery Tools Forum 2008
Outline

- Deep indexing
- Challenges
- Is there a need?
- Implementations
- Opportunities and dangers
- Issues for research and development
Deep Indexing

• Working definition:
  – An indexing system that supports discovery of information objects at levels of granularity beyond the abstract or article

• E.g., Figures, tables, other components drawn from scholarly journal articles
Deep Indexing

- Is there a need?
  - T&F hold data, images - organisms, objects, phenomena - in a condensed fashion
  - T&F important for judging article level relevance (Bishop, 1999; Sandusky & Tenopir, in press; Olsen, 1994; Stewart, 1996)
  - Users do seek photographs, maps, data with high specificity for teaching and research (Sandusky & Tenopir; Sandusky, Tenopir & Casado)
  - User want more efficient searching
    - When seeking a map or image, be able to find one
General Model & Challenges
Challenges

• Extracting & processing components of interest
• Indexing the components
• End-user experience
Challenges: Extraction & Processing

- Two extraction approaches
  - Identify, extract, and mark up components from untagged documents (e.g., PDF, Word)
  - Obtain tagged components from commercial / association publication processes

- Processing depends upon many factors
  - What do you want to expose?
  - How do you expose it?
Challenges: Indexing

• Computational indexing approaches
• Human indexing

• Choose elements to index:
  – Captions / legends
  – Content of tables
  – Text referring to tables and figures
  – Images
Challenges: End-User Experience

• End-user experience
  – Will users recognize they have direct access to components?
    • Will users know how to utilize it? (Bishop)
  – What does it mean to rank components returned in a search set? (Liu, et. al.)
  – What is a good component search interface? (Hearst, et al.)
Deep Indexing Implementations

• Four implementations
  – DeLLiver (Desktop Link to Virtual Engineering Resources) - 1996-2000
  – Illustrata - 2006+
  – TableSeer (CiteSeer) - 2006+
  – BioText Search Engine - 2007+
Deep Indexing Implementations

- DeLLiver (Desktop Link to Virtual Engineering Resources) - 1996-2000
  - Full text of 50 journals in engineering, physics, and computer science
  - SGML feeds directly from publishers, normalized and indexed
Deep Indexing Implementations

• DeLLiver
  – Direct, automatic indexing of components
    • Full text, abstract, title, author, author affiliation, section headings, journal title
    • Article body only
    • Cited references
    • Table captions
    • Table text
    • Figure caption
Deep Indexing Implementations

• DeLlver
  – “Extended citation”
    • Citation, affiliation, abstract, link to corresponding A&I index records
    • Link to full text
    • List of and links to figures and tables from article
      – Caption and icon (not image thumbnail)
    • References (with links)
    • Links to citing articles in database

FULL TEXT: XML/HTML APS PDF
CITATIONS & OTHER LINKS TO FULL TEXT: APS Abstract INSPEC

AUTHORS:

Maier, Robert S.
Department of Mathematics, University of Arizona, Tucson, Arizona 85721

Stein, D. L.
Department of Physics, University of Arizona, Tucson, Arizona 85721

ABSTRACT:

Suppose a two-dimensional dynamical system ...

© 1996 The American Physical Society 0031-9007/96/77(24)/4860(4)/$6.00

REFERENCES:


P. Haegg, P. Talkner, and M. Bockoven, Rev. Mod. Phys. 62, 251 (1990) INSPEC APS Abstract


FIGURES & TABLES:

FIG. 1. The unstable limit cycle $6W$ of the van der Pol model, and the MPEF, which emerges from the attractor $0,0$ and spirals into it. The trajectories exiting from $W$ are optimal trajectories that are small perturbations of the MPEF.

FIG. 2. A Poincare section. This sketch shows the points $p, q, n$ generated by the optimal trajectories passing by some specified point on $6W$. The dots are generated by the MPEF, spiraling into $6W$. Cf. Figs. 1-3 of Graham and Tel.

CITING ARTICLES:


Deep Indexing Implementations

• **Illustrata - 2006+**
  – Developed in 2005-2006 by CSA (now ProQuest CSA)
  – Prototype evaluated summer 2006
    • 300,000 figure / table objects
  – 60 scientists in nine universities / research institutes in U.S. and Europe
  – Product in 2007
    • 1,000,000 figure / table objects, most from 2000+
    • 1,000 peer reviewed journals
Deep Indexing Implementations

• Illustrata component search fields include
  – Object Descriptors
  – Object Geographic Terms
  – Object Statistical Terms
  – Object Subject Terms
  – Object Taxonomic Terms
  – Caption
1. Effect of ambient oxygen concentration on activities of enzymatic antioxidant defences and aerobic metabolism in the hydrothermal vent worm, Paralvinella gracile. Marine Biology [Mar. Biol.]. Vol. 160, no. 2, pp. 273-284. Nov 2006. gracile is a common endemic polychaete from the deep-sea hydrothermal vent communities located on the East Pacific Rise (EPR). These organisms colonise a large range of micronhabitats around active sites where physico-chemical conditions are thought ... View Record | FindText

2. Morphology, reproductive biology and genetic structure of the whale-fall and hydrothermal vent specialist, Bathylzkula guaymasensis (Annelida; Polynoidae) Glover, Adrian G; Gageza, Eliisa; Dahlgren, Thomas G; Smith, Craig R. Marine Ecology [Mar. Ecol.]. Vol. 26, no. 3-4, pp. 223-234. Sep 2005. Polynoidae, a species recorded from both whale-falls and hydrothermal vents. The habitat of B. guaymasensis is quite different from other vent polynoids, being apparently a specialist feeder on Beggiaota mat, rather than living commensally ...

3. An obligately photosynthetic bacterial anaerobe from a deep-sea hydrothermal vent Beauty, J Thomas; Overmann, Joerg; Linne, Michael T; Manske, Ann K; Lang, Andrew S; Blankenship, Robert E; Van Dover, Cindy L; Madinov, Tracy A; Plumley, F Gerald Proceedings of the National Academy of Sciences, USA [Proc. Natl. Acad. Sci. USA]. Vol. 102, no. 26, pp. 9306-9310. Jun 2005. ... of a previously unknown green sulfur bacterial species from a deep-sea hydrothermal
Enhanced abstract

We examined the reproductive biology and genetic structure of the polychaete Bathyporeia guaymasensis (Annelida: Polynoidae), a species recorded from both whale-fall and hydrothermal vents. The habitat of B. guaymasensis is quite different from other vent polychaetes, being apparently a specialist feeder on Beggiatoa mat, rather than living commensally with mussels or on the sides of hot vent chimneys. Specimens were collected from whale-fall sites in the Santa Cruz and Santa Catalina Basins, Southern California at depths of 1200-1600 m, and investigated using a combined morphological and molecular approach. The species exhibits marked sexual dimorphism, asynchronous gametogenesis, evidence for internal fertilization and lecithotrophic larval development - all characters shared with other hydrothermal vent polychaetes. Two distinct, sympatrically distributed haplotype groups of B. guaymasensis were identified using mitochondrial cytochrome oxidase 1 gene sequences, indicating that there may be two species within the nominal designation B. guaymasensis. Broad sharing of haplotypes between the two whale-fall sites suggests high dispersal rates among basins along the California margin.

Descriptors

- Article Subject Terms: Hydrothermal vents, Basins, Genetic structure, Haplotypes, Gametogenesis, Cytochrome-c oxidase, Fertilization, Mitochondria, Hydrothermal springs, Population genetics, Reproduction, Zoobenthos, Sexual dimorphism, Marine invertebrates, Deep water, Animal physiology, Animal morphology, Larval development
- Article Taxonomic Terms: Polynoidae, Annelida, Beggiatoa, Bathyporeia guaymasensis
- Article Geographic Terms: USA, California, USA, California, Santa Cruz, INE, USA, California, Santa Catalina Basin

Object Descriptors

- Object Subject Terms: Acicular chaetae, Cytochrome oxidase 1 haplotypes, Developing sperm, Enlarged nephridial...
Isotopic dating of the migration of a low-grade metamorphic front during orogenesis

Rasmussen, B., Fletcher, I. R., & Sheppard, S. (2005). Isotopic dating of the migration of a low-grade metamorphic front during orogenesis [Figure 1]. Geology, 33, 773-776.

Caption: Figure 1. A: Regional map of Pilbara Craton showing location of geochronology sample sites and U-Pb monazite and xenotime dates (quoted with 2 sigma precision). AB-Asburton Basin; BB-Bangemall Basin (recently divided into Collie and Edmund Basins); OB-Officer Basin. GC-Gascoyne Complex. Unconformity between Yuree Creek Group and Lower Wyloo Group spans duration of Chtehallian orogeny. B: Stratigraphic column and U-Pb zircon dates. Compiled from Pickard (2002), Buick et al. (1995, 2002), Trendell et al. (1998, 2004), and Mueller et al. (2005). FM-Formation, Gl-Group, Mt-Mount, Qtz-Quartzite, Spr-Spring-Springs, Suc-Succession. C: U-Pb ages of metamorphic monazite and xenotime showing younging trend from southwest to northeast.

Category: Figure; Graph; Scatter Plot; Map

Title: Isotopic dating of the migration of a low-grade metamorphic front during orogenesis

Author: Rasmussen, Birger; Fletcher, Ian R.; Sheppard, Stephen


Object Descriptors: Gascoyne Complex; Monazite dates; U-Pb zircon dates; Xenotime dates

Object Geographic Terms: Australia, Western Australia, Pilbara Craton
Deep Indexing Implementations

• TableSeer (CiteSeer) - 2006+
• Addressing:
  – Table search engine
  – Table metadata specification
  – Table detection / metadata extractor
    • Untagged documents
  – Table results ranking algorithm

• Not focused on interface
  – But see http://chemxseer.ist.psu.edu/
Deep Indexing Implementations

• BioText Search Engine - 2007+
  – 80,000 objects
  – 20,000 articles
  – 150 open access journals from PubMed Central
  – Growth in collection size & function
The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate


ABSTRACT

Ever since the discovery of 'genes in pieces' and mRNA splicing in eukaryotes, origin and evolution of spliceosomal introns have been considered within the conceptual framework of the 'introns early' versus 'introns late' debate. The 'introns early' hypothesis, which is closely linked to the so-called exon theory of gene evolution, posits that protein-coding genes were interrupted by numerous introns even at the earliest stages of life's evolution and that introns played a major role in the origin of proteins by facilitating recombination of sequences coding for small protein/peptide modules. Under this scenario, the absence of spliceosomal introns in prokaryotes... Show Full Abstract

FULL-TEXT EXCERPTS

...for bacteria than it is for organelles [57]. So what happened at the onset of eukaryotic evolution? The scenarios of eukaryogenesis favored by different researchers differ dramatically, and this is not the place to review in any detail the pros and cons of each of these scenarios [58]. In any case, there is no doubt that the symbiosis between an α-proteobacterium, the ancestor of the mitochondria, and a somewhat mysterious host, the ancestor of the eukaryotic... Show Full Excerpts

The ancient Virus World and evolution of cells


ABSTRACT

Recent advances in genomics of viruses and cellular life forms have greatly stimulated interest in the origins and evolution of viruses and, for the first time, offer an opportunity for a data-driven exploration of the deepest roots of viruses. Here we briefly review the current views of virus evolution and propose a new, coherent scenario that appears to be best compatible with comparative-genomic data and is naturally linked to models of cellular evolution that, from independent considerations, seem to be the most parsimonious among the existing ones. Several genes coding for key proteins involved in
The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?

ABSTRACT
Ever since the discovery of 'genes in pieces' and mRNA splicing in eukaryotes, origin and evolution of spliceosomal introns have been considered within the conceptual framework of the 'introns early' versus 'introns late' debate. The 'introns early' hypothesis, which is closely linked to the so-called exon theory of gene evolution, posits that protein-coding genes were interrupted by numerous introns even at the earliest stages of life's evolution and that introns played a major role in the origin of proteins by facilitating recombination of sequences coding for small protein/peptide modules. Under this scenario, the absence of spliceosomal introns in prokaryotes... Show Full Abstract

Figures From Article (3)

FIGURE CAPTION
Figure 1. Intron density in the genes of some modern eukaryotes and the reconstructed common ancestor of plants and animal (possible LECA under unicist-eocen phylogeny). The marks on the line show approximate intron density values (mean number of introns per gene); the...

FIGURE CAPTION
Figure 2. The proposed chain of causes and events in eukaryogenesis – the pivotal roles of mitochondrial endosymbiosis and intron invasion. Arrows indicate proposed causal relationships (selective forces).

FIGURE CAPTION
Figure 3. A brief early history of spliceosomal introns. The scheme shows the
Figure 1. Intron density in the genes of some modern eukaryotes and the reconstructed common ancestor of plants and animal (possible LECA under unicont-bicont phylogeny). The marks on the line show approximate intron density values (mean number of introns per gene); the data is from [99]. The reconstructed values for the ancestral genome (in red) are denoted after the reconstruction method: MCMC, Markov Chain Monte Carlo[49]; MP, maximum parsimony[23]; ML, maximum likelihood[41]; ML/MP, hybrid maximum likelihood/parsimony method[51].
The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate


FIGURE CAPTION
Figure 2. The proposed chain of causes and events in eukaryogenesis – the pivotal roles of mitochondrial endosymbiosis and intron invasion. Arrows indicate proposed causal relationships (selective forces).

VIEW FULL ARTICLE: HTML | PDF  (View all figures and tables from this article)

The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate


FIGURE CAPTION
Figure 1. Intron density in the genes of some modern eukaryotes and the reconstructed common ancestor of plants and animal (possible LECA under unicont-bicont phylogeny). The marks on the line show approximate intron density values (mean number of introns per gene); the data is from [99]. The reconstructed values for the ancestral genome (in red) are denoted... Show Full Caption

VIEW FULL ARTICLE: HTML | PDF  (View all figures and tables from this article)
Object search: results as grid

Figure 1. Infectious scenarios and treatment window for patients coinfected with hepatitis B and...
Host-virus interaction: a new role for microRNAs

<table>
<thead>
<tr>
<th>Source Virus</th>
<th>Virus Type</th>
<th>Number of microRNAs</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epstein Barr virus</td>
<td>Herpesvirus</td>
<td>32</td>
<td>[42]</td>
</tr>
<tr>
<td>Kaposi sarcoma-associated herpesvirus</td>
<td>Herpesvirus</td>
<td>17</td>
<td>[43]</td>
</tr>
<tr>
<td>Mouse gammaherpesvirus</td>
<td>Herpesvirus</td>
<td>10</td>
<td>[42]</td>
</tr>
<tr>
<td>Human cytomegalovirus</td>
<td>Herpesvirus</td>
<td>14</td>
<td>[42]</td>
</tr>
<tr>
<td>Herpes Simplex-1</td>
<td>Herpesvirus</td>
<td>1</td>
<td>[44]</td>
</tr>
<tr>
<td>Rhesus lymphocryptovirus</td>
<td>Herpesvirus</td>
<td>22</td>
<td>[45]</td>
</tr>
<tr>
<td>Simian virus 40</td>
<td>Papaviruses</td>
<td>2</td>
<td>[41]</td>
</tr>
</tbody>
</table>

The ancient Virus World and evolution of cells

<table>
<thead>
<tr>
<th>Concept</th>
<th>Principal message</th>
<th>References</th>
<th>Brief critique/comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloaca degeneration</td>
<td>Viruses at least complex ones evolved as a</td>
<td></td>
<td>This route of virus evolution appears to be inconsistent with</td>
</tr>
</tbody>
</table>
Interface Comparison

• BioText and Illustrata
• “Enhanced abstract” view of same article
The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?


Abstract

Ever since the discovery of 'genes in pieces' and mRNA splicing in eukaryotes, origin and evolution of spliceosomal introns have been considered within the conceptual framework of the 'introns early' versus 'introns late' debate. The 'introns early' hypothesis, which is closely linked to the so-called exon theory of gene evolution, posits that protein-coding genes were interrupted by numerous introns even at the earliest stages of life's evolution and that introns played a major role in the origin of proteins by facilitating recombination of sequences coding for small protein/peptide modules. Under this scenario, the absence of spliceosomal introns in prokaryotes... Show Full Abstract

View Full Article: HTML | PDF

Figures From Article (3)

**Figure Caption**

Figure 1. Intron density in the genes of some modern eukaryotes and the reconstructed common ancestor of plants and animal (possible LECA under unicist-eclont phylogeny). The marks on the line show approximate intron density values (mean number of introns per gene); the... Show Full Caption

**Figure Caption**

Figure 2. The proposed chain of causes and events in eukaryogenesis - the pivotal roles of mitochondrial endosymbiosis and intron invasion. Arrows indicate proposed causal relationships (selective forces).

**Figure Caption**

Figure 3. A brief early history of spliceosomal introns. The scheme shows the...
The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?

Koonin, Eugene V

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA


Background Ever since the discovery of “genes in pieces” and mRNA splicing in eukaryotes, origin and evolution of splicingosomal introns have been considered within the conceptual framework of the introns early versus “introns late” debate. The introns early hypothesis, which is closely linked to the so-called exon theory of gene evolution, posits that protein-coding genes were interrupted by numerous introns even at the earliest stages of life’s evolution and that introns played a major role in the origin of proteins by facilitating recombination of sequence coding for small protein/polypeptide modules. Under this scenario, the absence of splicingosomal introns in prokaryotes is considered to be a result of “genome streamlining”. The introns late hypothesis counters that splicingosomal introns emerged only in eukaryotes, and moreover, have been inserted into protein-coding genes continuously throughout the
Opportunities and dangers

• End-user evaluation (DeLiVer; Illustrata):
  – Higher precisions searching
  – F & T help relevance judgments
  – Direct access to high-quality artifact
    • Without getting paper
  – F & T use in teaching / presentation
Opportunities and dangers

• End-user evaluation (DeLiVer; Illustrata):
  – Disaggregated components are re-used in new text and performative documents
    • Papers, proposals, presentations
  – Support comparisons of many kinds
    • Sanity checks; find useful data (tables)
  – Creating data compilations
  – Finding data is cheaper than generating it
Opportunities and dangers

• However
  – Interface complexity
    • Filters (e.g., categories; dates) important
    • Illustrata vs. BioText
  – Broader communications impacts
    • Effect of decontextualization of components
    • Citing components? How and what creator benefit?
    • Preparing components that can stand alone
  – Information professionals
    • Promoting and supporting new discovery systems
    • Understanding disciplinary differences
Opportunities and dangers

• However
  – Incommensurate implementations
  – Lack of standards
  – Lack of openness
    • Proprietary system has greatest functionality
Issues for research and development

• The road to universal discovery
  – Metadata standards for independently findable, first-class objects / components
    • Tables, figures, maps, photographs, graphs now
    • What about
      – Datasets
      – Workflows
      – Simulations
      – Digitized specimens
      – Others?

  – Integration of component / object search and retrieval into next generation discovery tools
Issues for research and development

• Open extraction and indexing systems
  – Precision / recall test for technical evaluation
  – Need comparative, contextualized user studies (e.g., compare BioText to Illustrata)

<table>
<thead>
<tr>
<th>System</th>
<th>End Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeLLiver</td>
<td>108 + 350 surveys</td>
</tr>
<tr>
<td>Illustrata</td>
<td>60</td>
</tr>
<tr>
<td>BioText</td>
<td>8</td>
</tr>
<tr>
<td>TableSeer</td>
<td>0</td>
</tr>
</tbody>
</table>
Issues for research and development

• Between-discipline studies
  – Patterns of component discovery and use
    • Object types
    • Discovery and use patterns

• Improved contextualization
  – DeLiVer evaluation as model
  – Component systems as part of an ecology
    • BioText and Illustrata
Issues for research and development

• It’s a question of end-user benefits & costs:
  – Does focused searching on components yield improved results (better, faster)?
  – Does component availability transform scientific practice in positive ways?

OR

  – Are these additional separate systems to search?
  – Does component searching create larger results sets to evaluate?
Issues for research and development

• Practice transformations: writing & publishing to ease components use
  – Create objects as if they will stand alone
  – Publishing processes support automatic extraction and downstream processing
  – Use persistent object identifiers to cross-link
    • Articles
    • Components
    • Datasets
Sources

DeLlver


Illustrata


Sources

TableSeer (CiteSeer)


BioText Search Engine


Contact

Dr. Robert J. Sandusky
Assistant University Librarian for Information Technology
Clinical Associate Professor
Richard J. Daley Library
University of Illinois at Chicago
sandusky@uic.edu