Smart searching on academic content platforms

Wouter van der Velde – eProduct manager eBooks
Merci....et, je m’excuse...

- Merci a vous...d’écouter à moi!
- Je m’excuse que parlez Français est difficile pour moi.
- La presentation est en Anglais
Who is Springer?

- Leading global scientific publisher
- 6,000 employees in 25 countries
- 890 million EUR in turnover
- 2,000 journals / 7,000 new book titles published every year
- 50,000 eBooks
- Largest open access portfolio worldwide (over 300 open access journals)
Sure, we’ve got what you want. In here somewhere.
Before I’ll show some smart search implementations...

• I’ll give you some background on what you do not see....
• But, what is part of each document
• And necessary to make recommendation and smart searching possible
From Raw Material towards Products and Services

Adding Value

- Raw Material
  - Intellectual Content
    - Reviewed by Peers
  - Manuscript
  - Images
  - Datasets
  - Multimedia Files

- Refined & Standardized
  - Typesetting
  - Copy Editing
  - Artwork
  - XML Creation
  - Proofed by Author

- Metadata Added
  - Authors, Title
  - ISSN, ISBN, DOI
  - Publication Date
  - Prices, Rights

- Shippable Product
  - Compiling
  - Printing
  - Binding
  - Packaging

- Availability & Findability
  - Online Archive
  - Digital Preservation
  - Citable, Unchanging and Persistent
  - DOI & CrossRef Reg.

- Context Added
  - Linked References
  - Forward Linking
  - Related Publication
  - Semantic Linking
Fly stocks

The stock y w sn^3 otu''; SuUR^+ carries the otu mutation (which is responsible for polytene chromosome formation in NCs) and two wild-type alleles of the SuUR gene. The stock sn^3 otu''; SuUR carries the otu and SuUR mutations in homozygous state. The stock y w sn^3 otu''; P[w' SuUR^+] carries the otu mutation and is homozygous for the insertion of a P-element-based construct X6S1 on chromosome 2. This construct contains the mini-white gene and the full-length genomic fragment of the SuUR gene (Mehnert et al. 2003). Thus, this stock carries four copies...
XML – Extensible Markup Language

Metadata

- Lingua Franca of the internet. Machine readable information
- Digital distribution: Third party, Libraries, A&I, Google, Amazon
- Content without correct metadata is worthless in the digital world

Full-text

- Cost and time saving in production, e.g. automated pagination
- Media neutral: Re-use, Re-purpose, Re-package
- Long term preservation (open standard, independent of application)

References

- Reference Linking, Forward Linking
Linking up the content in different ways

1. **Related Articles (Fingerprinting)**
   Showing the user of an SpringerLink article the 10 most closely related documents on SpringerLink

2. **Background-Enriched Content (BEC)**
   Enrich the full-text of all SpringerLink documents in html format with hyperlinks to related content (in development)
Related Articles

Method

1. Calculate fingerprints for each article and chapter
2. Compare fingerprints of all articles with one another
3. Identify the 10 most closely related articles for each article
Related Articles by Fingerprints on SpringerLink

Total: 5.7 million
Journal articles: 4.7 million

Example:
10.1007/s00261-010-9615-7

Acute gastrointestinal bleeding: CT angiography with multi-planar reformatting
Klaus Siewert, Frank Gerhard, Steffen Studer, Annette Poppa, Richard Wiedenheft
Department of Surgery, University Hospital, D-53105 Bonn, Germany

Abstract
Acute gastrointestinal bleeding is a common medical emergency that often requires CT angiography. To achieve a high level of diagnostic accuracy, it is essential to perform a multi-planar reformatting of the CT images. The aim of this study was to evaluate the diagnostic accuracy of CT angiography with multi-planar reformatting in patients with acute gastrointestinal bleeding.

Table:

<table>
<thead>
<tr>
<th>Term</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>cta image</td>
<td>103.89</td>
</tr>
<tr>
<td>axial cta image</td>
<td>95.55</td>
</tr>
<tr>
<td>contrast extravasation</td>
<td>68.19</td>
</tr>
<tr>
<td>bleed point</td>
<td>62.34</td>
</tr>
<tr>
<td>important non-invasive tool</td>
<td>50.63</td>
</tr>
<tr>
<td>b1 coronal</td>
<td>49.38</td>
</tr>
<tr>
<td>sma injection</td>
<td>43.98</td>
</tr>
<tr>
<td>acute gastrointestinal bleed</td>
<td>40.36</td>
</tr>
<tr>
<td>subsequent endovascular</td>
<td>37.95</td>
</tr>
<tr>
<td>mips demonstrate active extravasation</td>
<td>36.47</td>
</tr>
<tr>
<td>mips demonstrate active extravasation of</td>
<td>36.47</td>
</tr>
<tr>
<td>upper gi endoscopy</td>
<td>35.35</td>
</tr>
<tr>
<td>bleed source</td>
<td>35.77</td>
</tr>
<tr>
<td>gi endoscopy</td>
<td>34.82</td>
</tr>
<tr>
<td>pancreatic [pancreatic]</td>
<td>31.86</td>
</tr>
<tr>
<td>common medical emergency</td>
<td>25.33</td>
</tr>
<tr>
<td>bowel lumen</td>
<td>25.14</td>
</tr>
<tr>
<td>upper gastrointestinal bleed</td>
<td>24.72</td>
</tr>
<tr>
<td>attenuation fluid</td>
<td>24.54</td>
</tr>
<tr>
<td>ligament of treitz</td>
<td>24.46</td>
</tr>
<tr>
<td>rare cause of gastrointestinal bleed</td>
<td>24.08</td>
</tr>
<tr>
<td>active bleed</td>
<td>23.70</td>
</tr>
<tr>
<td>low gastrointestinal hemorrhage</td>
<td>22.90</td>
</tr>
<tr>
<td>smw varix</td>
<td>22.81</td>
</tr>
<tr>
<td>gastrointestinal hemorrhage</td>
<td>21.80</td>
</tr>
<tr>
<td>small pseudoaneurysms</td>
<td>21.78</td>
</tr>
<tr>
<td>active extravasation of</td>
<td>21.59</td>
</tr>
<tr>
<td>focal site</td>
<td>19.15</td>
</tr>
<tr>
<td>upper gastrointestinal hemorrhage</td>
<td>18.55</td>
</tr>
<tr>
<td>active extravasation</td>
<td>18.22</td>
</tr>
<tr>
<td>acute upper gi bleed</td>
<td>17.64</td>
</tr>
<tr>
<td>varix</td>
<td>16.46</td>
</tr>
<tr>
<td>left gastric artery</td>
<td>16.38</td>
</tr>
<tr>
<td>significant mortality</td>
<td>16.20</td>
</tr>
<tr>
<td>active hemorrhage</td>
<td>15.87</td>
</tr>
</tbody>
</table>
Related Articles
Related Articles

Article Fingerprint → Mathematical Similarity Analysis → Rank Related Articles

with all Springer content

„Article-Article Correlation“
10 most Related documents on SpringerLink with every article
Background-Enriched Content

Oncogene

Manfred Schwab

1 DKFZ, Heidelberg, Germany

Email: m.schwab@dkfz.de

Without Abstract

Definition

An oncogene is a derivative of any gene that has the ability to stimulate cellular growth. In experimental assays, oncogene products can, alone or in cooperation with another gene, transform eukaryotic cells so that they grow in a way analogous to tumor cells. The definition was originally applied to the transforming genes acquired by RNA tumor viruses through the transduction of cellular genes. Today, the term is used rather broadly. Oncogenes contribute to tumorigenesis by any positive modulation of cellular growth; they act by their presence (this in contrast to tumor suppressor genes), an activity that is often referred to as “dominant”. Tumorigenic activation of oncogenes can result from mutational/structural/numeric changes in a gene and possibly from regulatory enhancement of gene expression.

Characteristics

Oncogenes were originally isolated from RNA tumor viruses, where they are responsible for the rapid tumor induction after infection of an animal host. In the viral genome, the oncogene was referred to as a viral oncogene or v-onc (1, 2).

It was soon established that the v-oncs are actually derived from the genome of the host cell. They have been captured by the virus after infection of the cell by a process called transduction. Transduction appears in a range of animal species from chickens to monkeys; it has not been observed in humans. The cellular counterparts, from which the v-oncs are derived, are referred to as proto-oncogenes, or cellular oncogenes (c-onc). Proto-oncogenes are normal constituents of the cellular genome and are highly conserved among all eukaryotic organisms.

This original rigid definition has softened in subsequent years. Broadly speaking, the term oncogene now includes any gene that has a growth stimulatory effect on cells, by means of:

- conferring sustained cellular multiplication
- advancement of cell-cycle progression
- decreased requirement for growth factors
- focus formation under conditions of cell culture
- enabling cells to grow under more restricted experimental conditions, such as in soft agar
- tumorigenic conversion, such as in experimental animals
Background-Enriched Content Method

1. Identify meaningful terms in full text of every HTML article
2. For every term, find articles relevant for the term...
3. ...and with matching fingerprints
4. Rank these articles and render the hit list upon clicking on the term
And....what about the usage?

“On SpringerLink, the usage of related articles is substantial. In Q1 2012, we had 17 million ‘related articles’ abstract views.”
SpringerLink for corporate customers...
Suggested Search

- We took a listing of over 900,000 keywords that drove traffic to SpringerLink and use that as the basis, and when you type, the autosuggest is triggered after the 3rd keystroke and shows you the matching results in order of how much traffic that term brought to our site.
Search customizations

• Future search enhancements:
  
  • Highly cited weigh more
  
  • Highly downloaded weigh more
Summary

• SpringerLink hosts 5.7 million articles and chapters
• Metadata quality is essential
• SpringerLink offers 10 most related documents
  — Based on fingerprint
• SpringerLink for Corporate has suggested search
  — Based on actual usage: give your user what they most likely need
• Researchers and students can find that needle in the haystack!
Thank You! – Questions?

Wouter van der Velde
  eProduct manager eBooks

Wouter.vanderVelde@springer.com