Journal Article Content Mining & Scholarly Publishers

STM Innovations Seminar; 2 December 2011

Maurits van der Graaf; Pleiade Management & Consultancy

Contents

The essence of content mining (incl. text mining and data mining):

If you have too much to read, or too much information to digest, could a machine do it for you?

Let's start with a potential user (1)

Use-case-1: keeping up-to-date

- Since 1982:
  - 90,000 journal articles on neuroregeneration (e.g. spinal cord injury)
- New articles:
  - on average 22 journal articles per day on neuroregeneration

Prof. Joost Verhaagen PhD, Netherlands Institute for Neuroscience, Amsterdam

Let's start with a potential user (2)

Use-case-2: Information needed as result of laboratory experiments

- Which molecules play a role in this process?
- Typical outcome of an experiment: hundreds of molecules show enhanced activity
- Next step: how to filter out the relevant molecules?
- *'One would like to have for each of these molecules a meta-analysis about what is already known about these molecules in other processes'*

Prof. Joost Verhaagen PhD, Netherlands Institute for Neuroscience, Amsterdam

INTRODUCTION PRC STUDY

Study commisioned by the Publishing Research Consortium

- Authors:
  - Eefke Smit, Bronfonteyn
  - Maurits van der Graaf, Pleiade Management & Consultancy
- Two parts:
  - Qualitative study:
    - 29 interviews with experts in academia, research, libraries, vendors and publishers
  - Quantitative study
    - Survey among publishers (members Crossref & STM)
    - 190 responses
- Full report on PRC website www.publishingresearch.net
- Article in the 1st issue of 2012 of Learned Publishing

Research questions

1. What is the present state of content mining?
2. What is the demand for content mining and how are publishers responding to this?
3. Which content mining opportunities are scholarly publishers pursuing in projects and plans ?
4. What are their expectations for the future?
5. What are possible cross-publisher solutions to facilitate journal article content mining better?

- Definition of content mining
- Developments in content mining
- Publisher’s dilemma of derivative products
- Optimists and pessimists

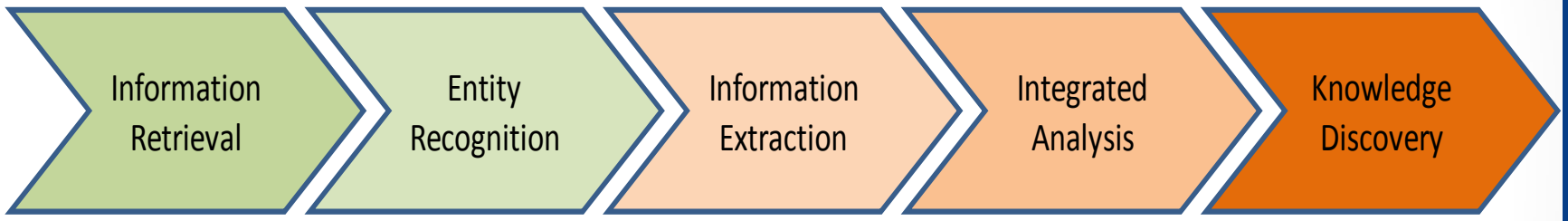
PRESENT STATE OF CONTENT MINING

Definition of content mining

- **Automated** tool or technique, or technology to **process large volumes** of digital textual content that is often **unstructured** or not uniformly structured
- **Purposes:**
  - To **identify** and select relevant information
  - To **extract** information from the content
  - To identify **relationships** within/ between/ across documents and between **incidents, events**, for meta-analysis

Developments in content mining

- Evolutionary stages of Content Mining:



Well established, used by A&I, search, annotation services

Content enrichment is hard coded (keywording, thesauri) or ad hoc, using document similarity, document clustering, subject categorization, mostly a combined approach.

Rapidly being improved

Searches for relationships, using co-occurrency methods with categorization techniques, needs disambiguation by curators

Infancy stage, promises for discovering new (cor)relations

Looks for worldwide trends across separate sources, historical or forecasting, seeks related objects and inference of overlooked relations, ambition to mine text and other formats alike.

Derivative information products

- Derivative information products that summarize and display the information and relationships as available in the content corpus that was mined (*information extraction*).
- Dilemma for publishers:
  - Positive usage effects: traffic drivers to content
  - Negative usage effects: function as substitute to original content

Optimists and Pessimists

Skeptics:

- Works only in very specialized, dedicated areas
- Tools not so automated, need a lot of work
- Manual curation still necessary
- High investments needed, business case difficult
- Domain dependent, needing good dictionaries
- Too complicated for an average user
- Will it ever fulfill its longstanding promise ?

Optimists:

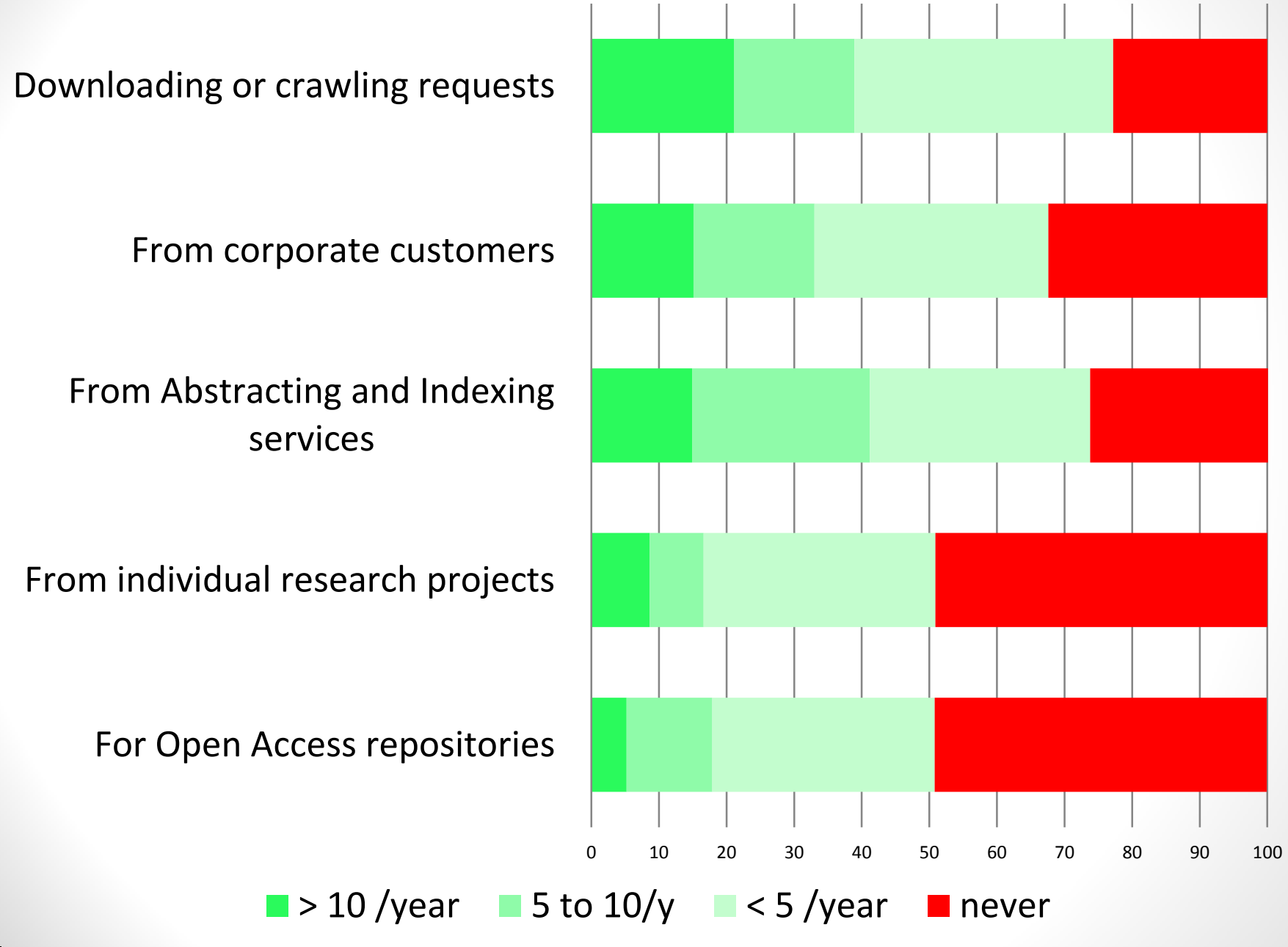
- Many broader applications appearing
- Tools are getting better and simpler
- For less precision, fully automated tools available
- Marketing and advertising have found good business cases
- For more general use, simple vocabularies suffice
- Mining will be an everyday tool
- The moment is now:

✓ Much more digital content available
✓ Computing capabilities and better tools
✓ Accessibility less of an issue

MINING BY THIRD PARTIES

- Frequency
- By whom?
- Increase
- Publishers' policies

Mining by third parties



Frequency of mining requests by 3rd parties

- **Requests wide-spread:**
  - 77% of the publisher respondents report such requests
  - Rest does not receive mining requests: small or OA publishers
- **Frequency rather low but increasing:**
  - Only 21% report more than 10 requests/year
  - Requests by: A&I services (74%); corporate customers (68%); research projects (51%)
  - 48% see increase in requests

Publishers' policies re mining requests by 3rd parties

- **32% without restrictions (most OA publishers)**
- **Rest considering requests case-by-case, balancing own interest and 3rd party interest:**
  - 85% require information about purpose in advance
  - 35%: permission in most/100% of the cases for all requests
  - 60%: permission in most/100% of the cases for research purpose requests
  - 60%: permission in most/100% of the cases for creation navigational products that drive more traffic to their content
  - 53%: permission declined for creation products that compete/replace their original content offerings
  - 31%: in most/100% cases a fee is asked in case of commercial purposes

MINING BY PUBLISHERS

- Present internal practices
- Future plans and intentions
- Expectations

46% of the publishers presently undertakes content mining on their own content:

we mine our content for the purpose of: (n=88)	
Improved information retrieval and better navigation to our content for example to create indexes, allocate keywords, to identify authors, affiliations etc.	81.8 %
For the generation of better metadata and for semantic tagging for example of genes, proteins, chemical compounds, diseases, people, places, organisations etc.	63.6 %
To create new products and services such as databases with factual information, visual representations, trends studies, performance information (citation analysis, usage statistics), RDF triple stores etc.	56.8 %
For our own business purposes such as competitive intelligence, market trends, research trends, customer reactions etc.	46.6 %

Plans within the next year re content mining on their own content

Improved information retrieval and better navigation to our content Plans within the next year on at least one of these topics: to improve our search engine, to create indexes, allocate keywords, to identify authors, affiliations etc., to add better browsing trails to related articles, to generate bibliographic information	48 %
For the generation of better metadata and for semantic tagging Plans within the next year or at least one of these topics: genes, proteins, chemical compounds, diseases, people, places, organisations, species, products, vendors	24 %
To create new products and services Plans within the next year or at least one of these topics: databases with factual information, visual representations, trends studies, performance information (citation analysis, usage statistics), topical bibliographies, RDF triple stores, search engines	36 %

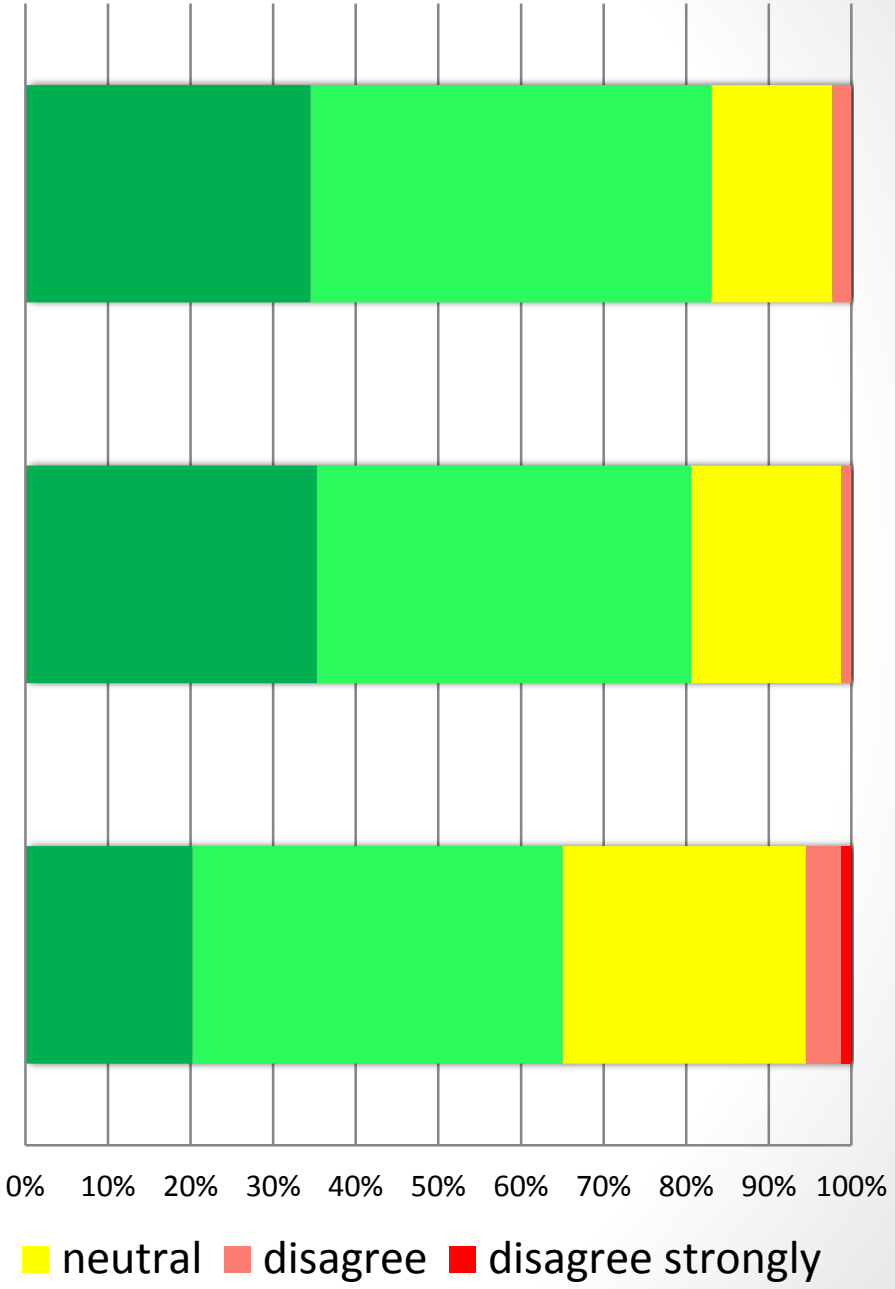
Of the 54% publisher respondents without existing internal practices re content mining, 36% report concrete plans for the next year

High expectations for content mining in publishing

Scholarly publishers will mine their content for the purpose of content enrichment, semantic tagging and better navigation.

Content mining will rapidly expand into new areas, new applications and new directions

More new services like Mendeley and Citeseer will emerge as a result of better content mining technology.

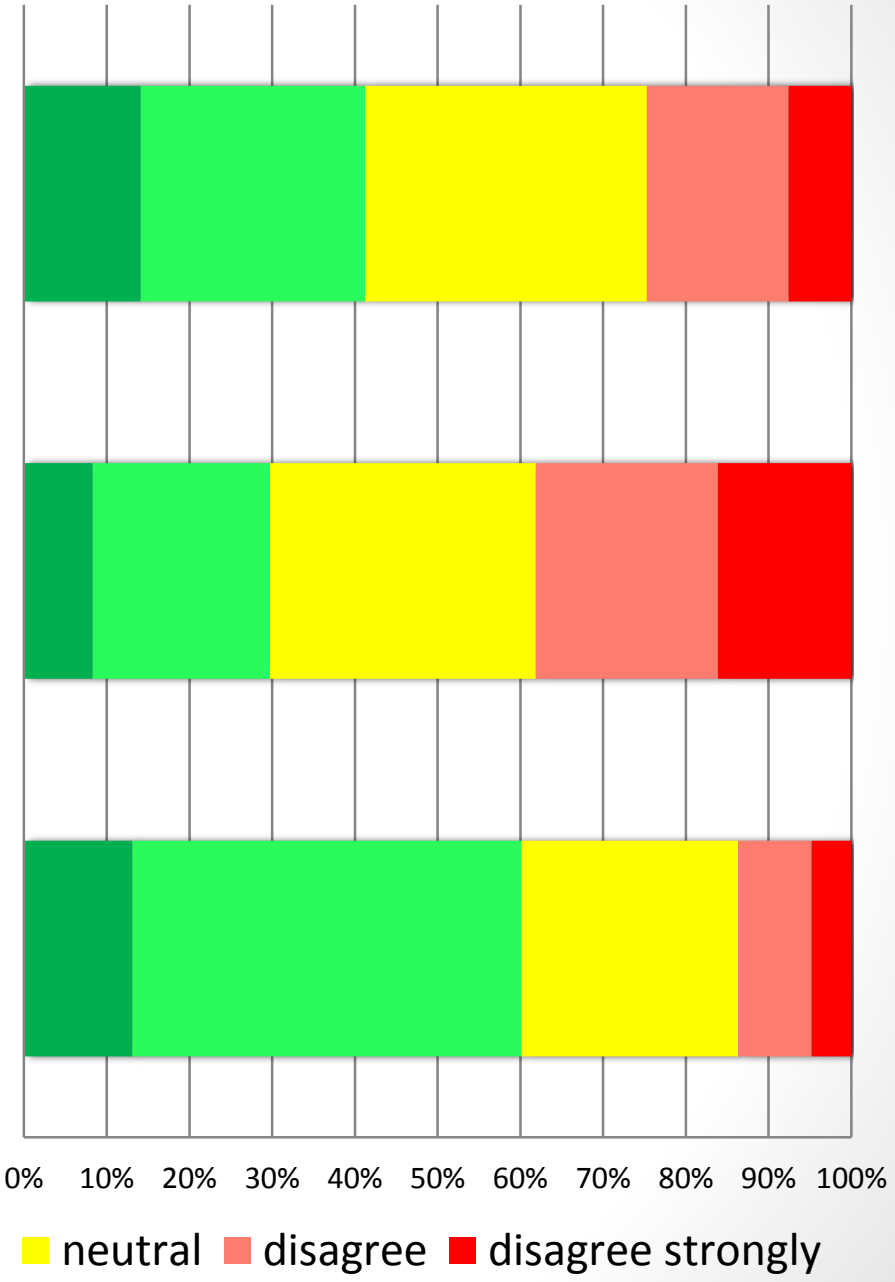


But differing expectations on:

Content mining tools will deliver high precision results in an automated way, without manual curation

The copyright protection for derivative works (created via mining for example) should be made less restrictive.

The investments needed for semantic tagging of scholarly content will be a limiting factor.



- Obstacles for content mining
- 5 solutions proposed by the interviewees

COMMON CROSS PUBLISHER SOLUTIONS

Obstacles for content mining

- The content spread over different publisher platforms
- The wide variety in permission rules for content mining
- Too much variety in content formats

Popularity of the proposed common, cross publisher solutions	Popularity among all respondents	Popularity among 'expert' respondents
Standardization of content formats for mining, of API-platform standards, of basic semantic tagging terms, etc.	52.8	67.5
One content mining platform	36.9	17.7
Commonly agreed access terms for content mining with a clear research focus and no commercial purpose	33.9	26.4
Central window to handle permission requests across publishers	27.5	17.8
Collaborate with national libraries to manage platforms for cross publisher content mining	9.5	-4.1

Conclusions

- Present state of content mining: most developments in information extraction (i.e. derivative products)
- Third party demand for content mining is widespread but (still) at low levels of frequency
- Publishers' permissions for content mining are quite liberal, especially for research-driven mining requests
- Approx. half of the publishers undertake mining of their own content
- Content mining is on the rise among publishers
- Most feasible cross publisher solutions:
  1. Standardisation of Content Formats
  2. Commonly agreed permission terms for research focused mining