stm

INTERNATIONAL ASSOCIATION OF SCIENTIFIC, TECHNICAL & MEDICAL PUBLISHERS

Content Mining,

a short introduction to practices and policies

Summary of
a study for the Publishing Research Consortium
into Journal Article Mining,
By Eefke Smit, Maurits van der Graaf (2011)

Full study available on PRC website

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

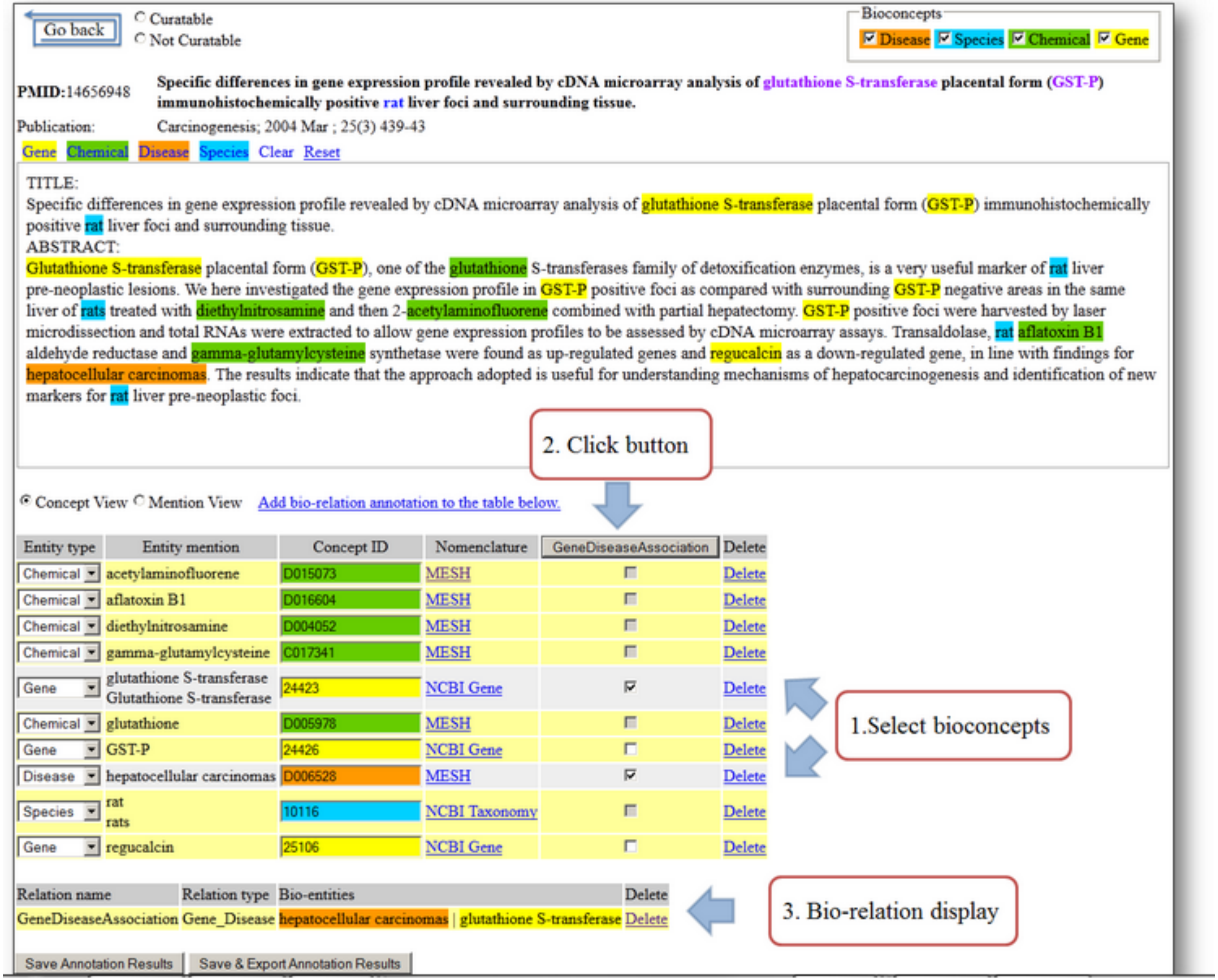
- Which molecules do play a role in this process?
- Typical outcome of an experiment: hundreds of molecules show enhanced activity
- Next step: how to filter out the relevant molecules?
- *‘You would like to have for each of these molecules a meta-analysis about what is already known about these molecules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

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

The essence of TDM is:

So much information to analyse:

Can a machine do this for him ?



Go back
○ Curatable
○ Not Curatable
Bioconcepts
☑ Disease ☑ Species ☑ Chemical ☑ Gene

PMID:14656948 **Specific differences in gene expression profile revealed by cDNA microarray analysis of glutathione S-transferase placental form (GST-P) immunohistochemically positive rat liver foci and surrounding tissue.**

Publication: Carcinogenesis; 2004 Mar ; 25(3) 439-43

Gene Chemical Disease Species Clear Reset

TITLE:
Specific differences in gene expression profile revealed by cDNA microarray analysis of glutathione S-transferase placental form (GST-P) immunohistochemically positive rat liver foci and surrounding tissue.

ABSTRACT:
Glutathione S-transferase placental form (GST-P), one of the glutathione S-transferases family of detoxification enzymes, is a very useful marker of rat liver pre-neoplastic lesions. We here investigated the gene expression profile in GST-P positive foci as compared with surrounding GST-P negative areas in the same liver of rats treated with diethylnitrosamine and then 2-acetylaminofluorene combined with partial hepatectomy. GST-P positive foci were harvested by laser microdissection and total RNAs were extracted to allow gene expression profiles to be assessed by cDNA microarray assays. Transaldolase, rat aflatoxin B1 aldehyde reductase and gamma-glutamylcysteine synthetase were found as up-regulated genes and regucalcin as a down-regulated gene, in line with findings for hepatocellular carcinomas. The results indicate that the approach adopted is useful for understanding mechanisms of hepatocarcinogenesis and identification of new markers for rat liver pre-neoplastic foci.

2. Click button

● Concept View ○ Mention View Add bio-relation annotation to the table below.

Entity type	Entity mention	Concept ID	Nomenclature	GeneDiseaseAssociation	Delete
Chemical	acetylaminofluorene	D015073	MESH	□	Delete
Chemical	aflatoxin B1	D016604	MESH	□	Delete
Chemical	diethylnitrosamine	D004052	MESH	□	Delete
Chemical	gamma-glutamylcysteine	C017341	MESH	□	Delete
Gene	glutathione S-transferase Glutathione S-transferase	24423	NCBI Gene	☑	Delete
Chemical	glutathione	D005978	MESH	□	Delete
Gene	GST-P	24426	NCBI Gene	□	Delete
Disease	hepatocellular carcinomas	D006528	MESH	☑	Delete
Species	rat rats	10116	NCBI Taxonomy	□	Delete
Gene	regucalcin	25106	NCBI Gene	□	Delete

1.Select bioconcepts

Relation name	Relation type	Bio-entities	Delete
GeneDiseaseAssociation	Gene_Disease	hepatocellular carcinomas | glutathione S-transferase	Delete

3. Bio-relation display

Text mining tool for semantic search by PubTator, see
http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/tutorial/index.html

Typical text mining consists of

- Processing large corpora of text in an automated way
- To identify entities, instances, actions, relationships and patterns and also for assertion- and sentiment analysis
- For example: genes, proteins, gene-disease patterns, compound properties, chemical structures, side effects of drugs
- **Text mining output typically consists of:**
  - Article clusters and categorisations, indexes
  - Topical maps, to show the occurrence of topics and their inter-relationships
  - Databases with facts, patterns, relationships, statements, assertions, properties found in the articles,
  - Visualisations like graphs, mappings, plot-graphs and topical maps

Optimists and Pessimists on TDM

Skeptics:

- Has always over-promised
- Only in specialized fields
- Tools still complicated
- Manual curation necessary
- High investments
- Domain dependent
- No common dictionary
- Overambition in the promise of knowledge discovery

Optimists:

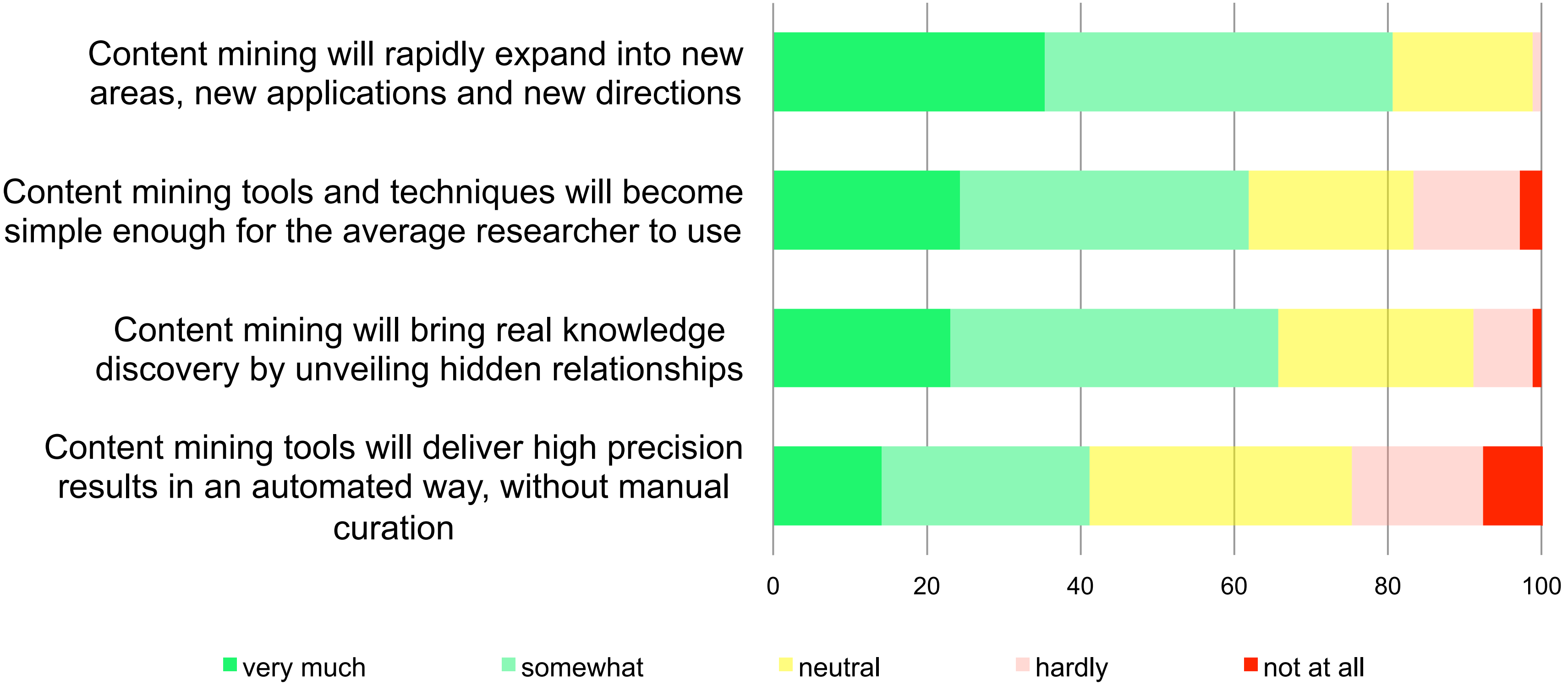
- Vast digital corpus available and growing
- More and more application areas (business, legal, social, etc)
- Tools improving fast
- Manual work reduced
- Public domain or domain precision
- Processing power less of a problem, analytical tools better, visualisation adds to analysis

Study commisioned by
the Publishing Research Consortium

- Authors:
  - Eefke Smit,
  - 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

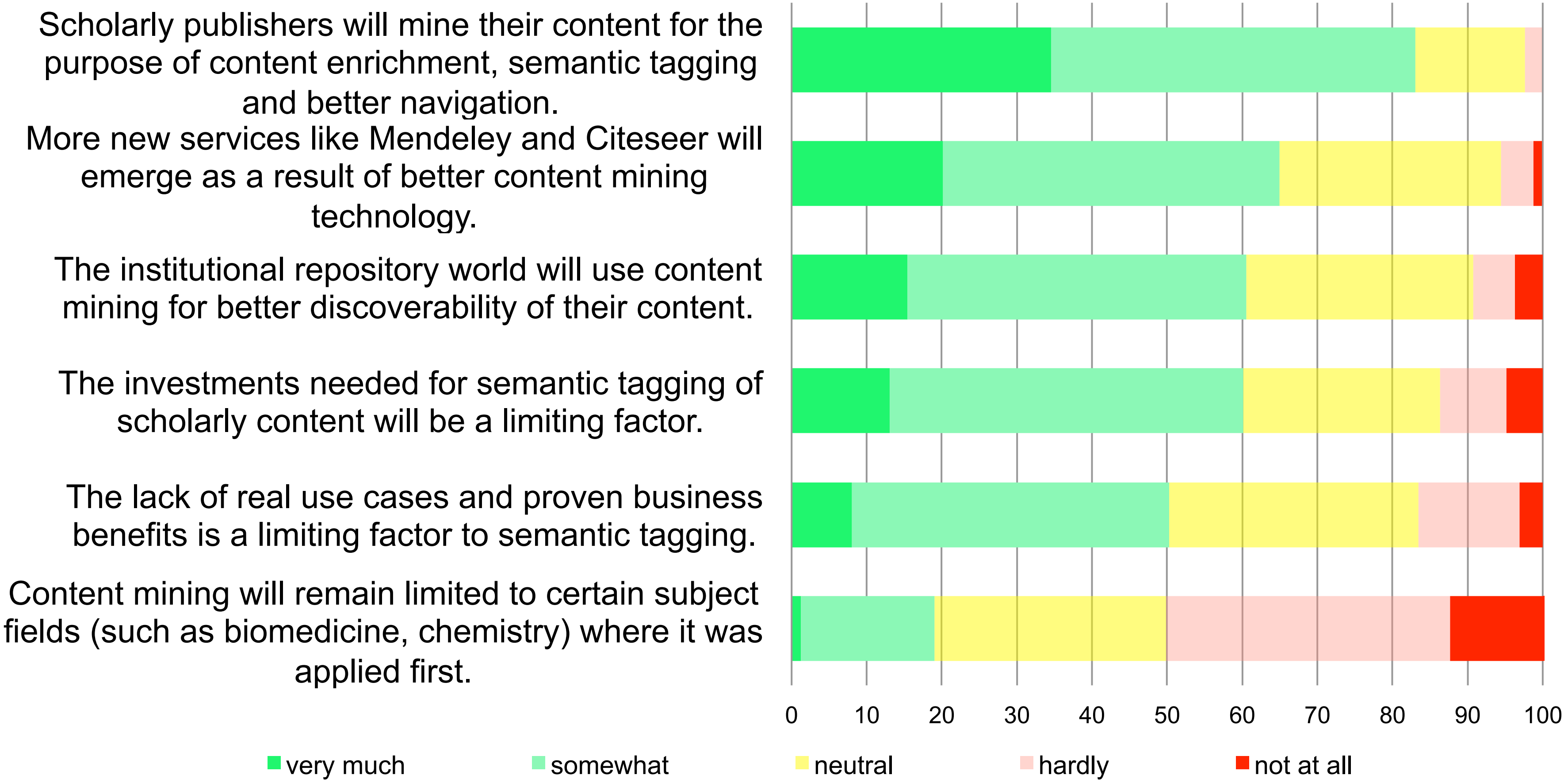
Publishers are optimistic:

Opinions/ expectations for Content Mining in the next 3 years

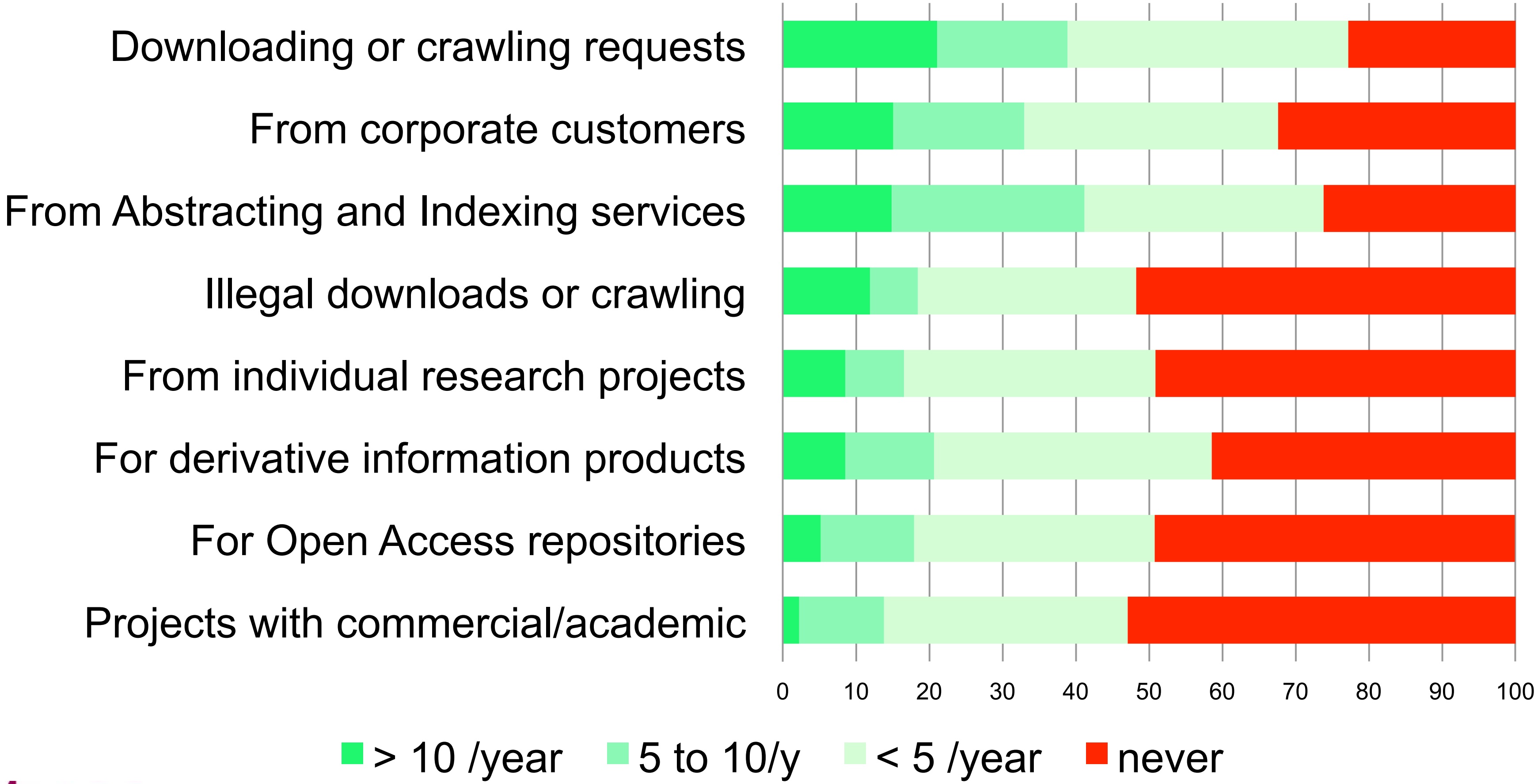


Publishers are optimistic, continued:

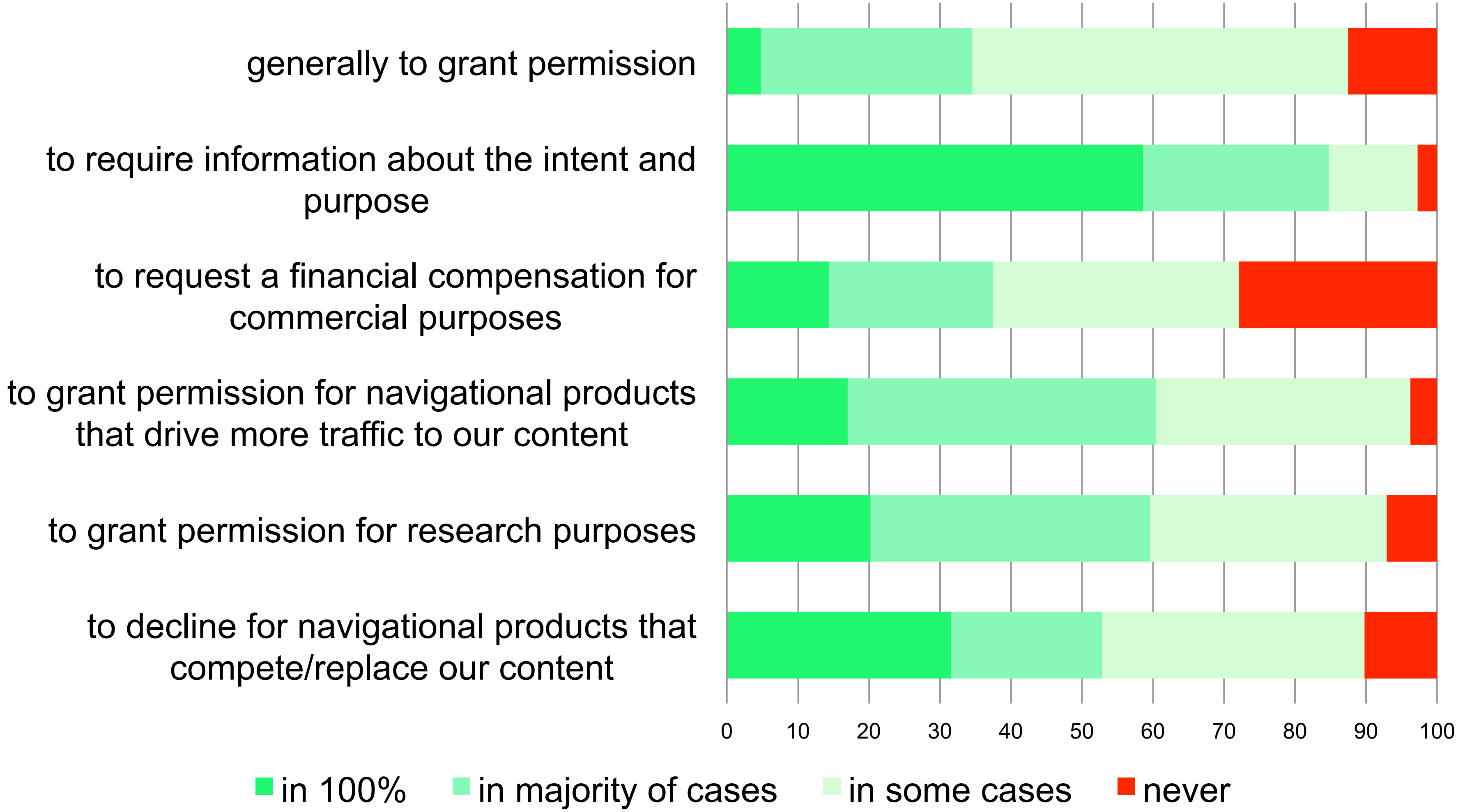
Opinions/ expectations for Content Mining on scholarly content in the next 3 years



…but publishers do not get many mining requests from 3rd parties:

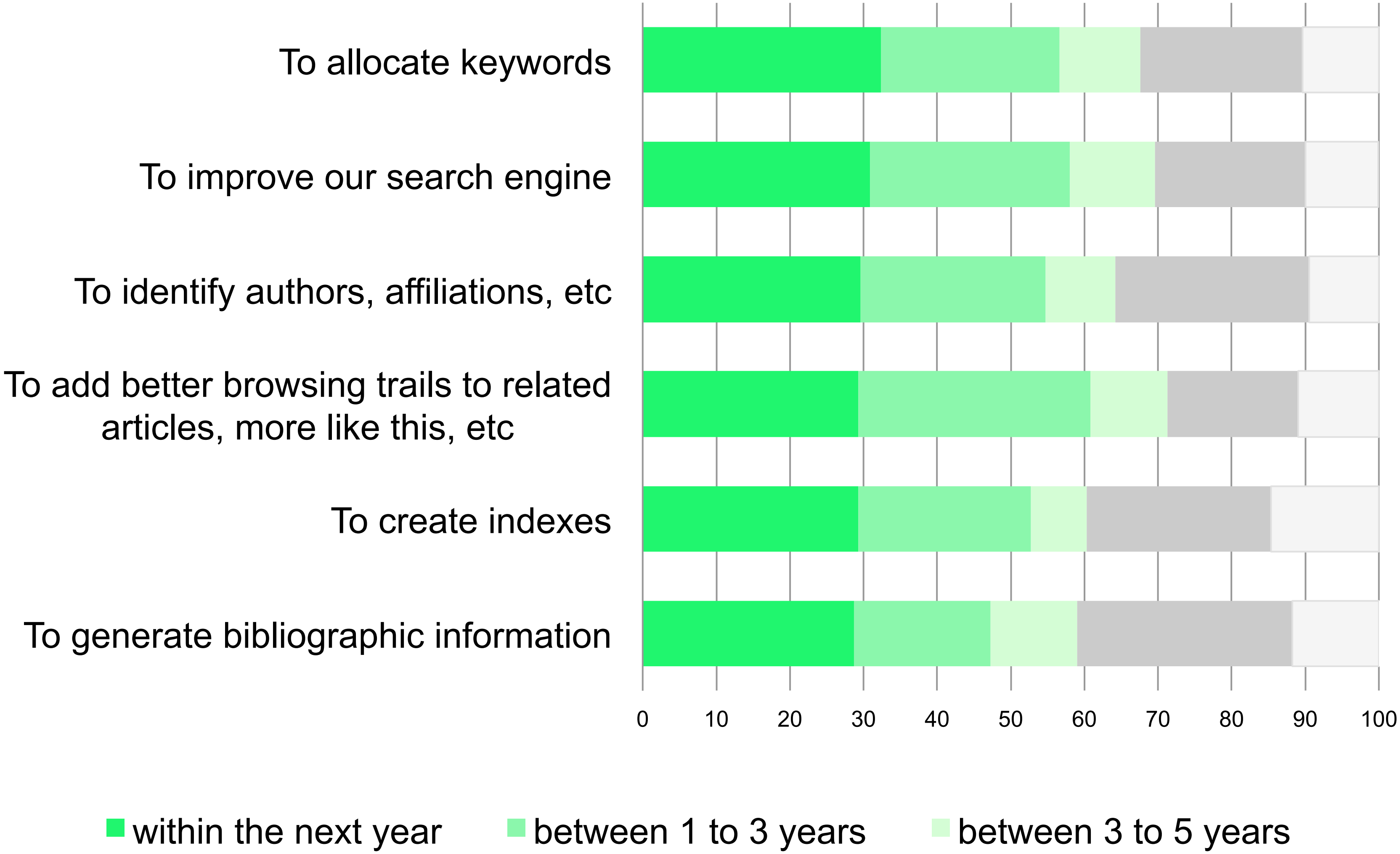


Publishers are liberal in allowing mining:
How case-by-case requests are treated



stm

…and plan more mining themselves: for retrieval and navigation



Cross-sector solutions to facilitate Content Mining better

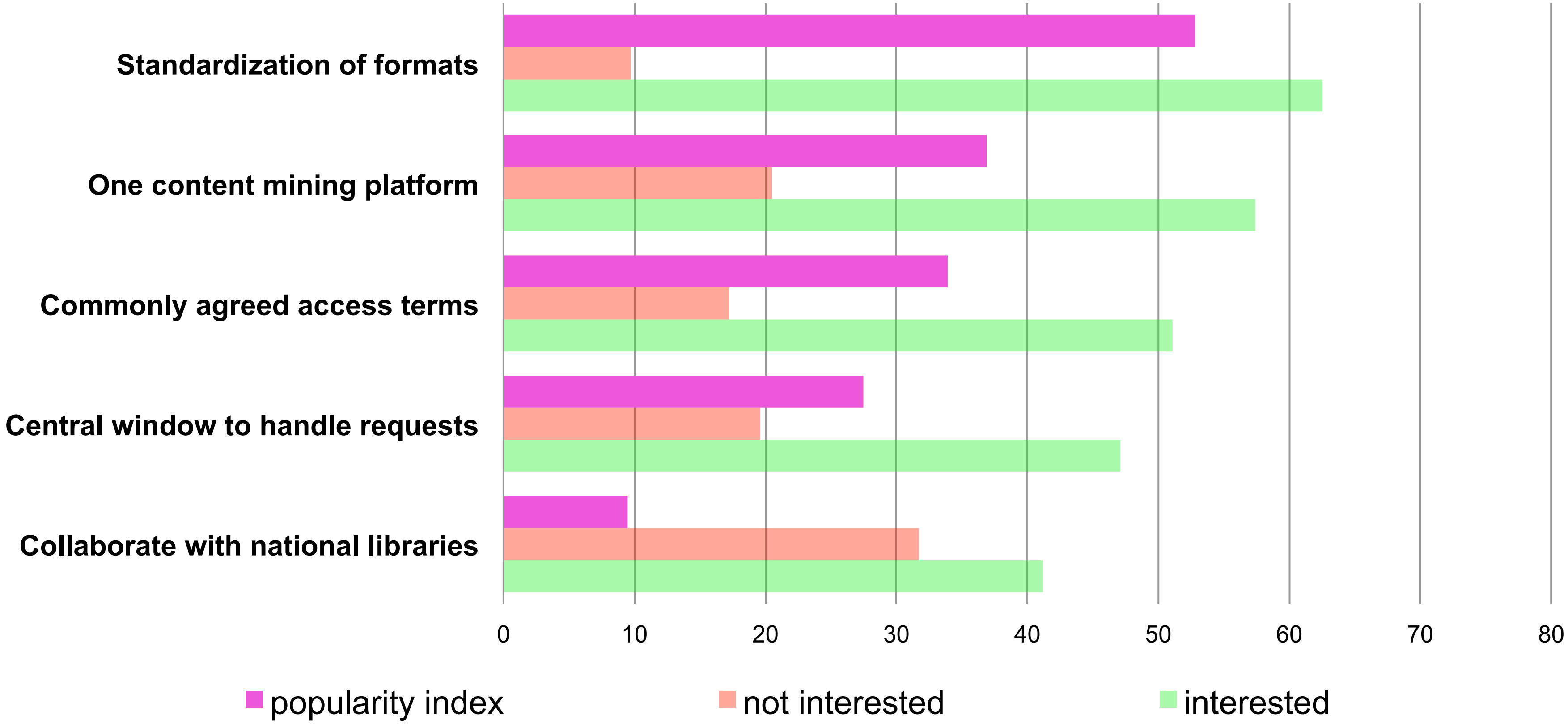
Suggestions made by experts during the interviews:

1. Standardization of Content Formats
2. One Content Mining platform
3. Commonly agreed access terms
4. One window for mining permissions
5. Collaboration with national libraries

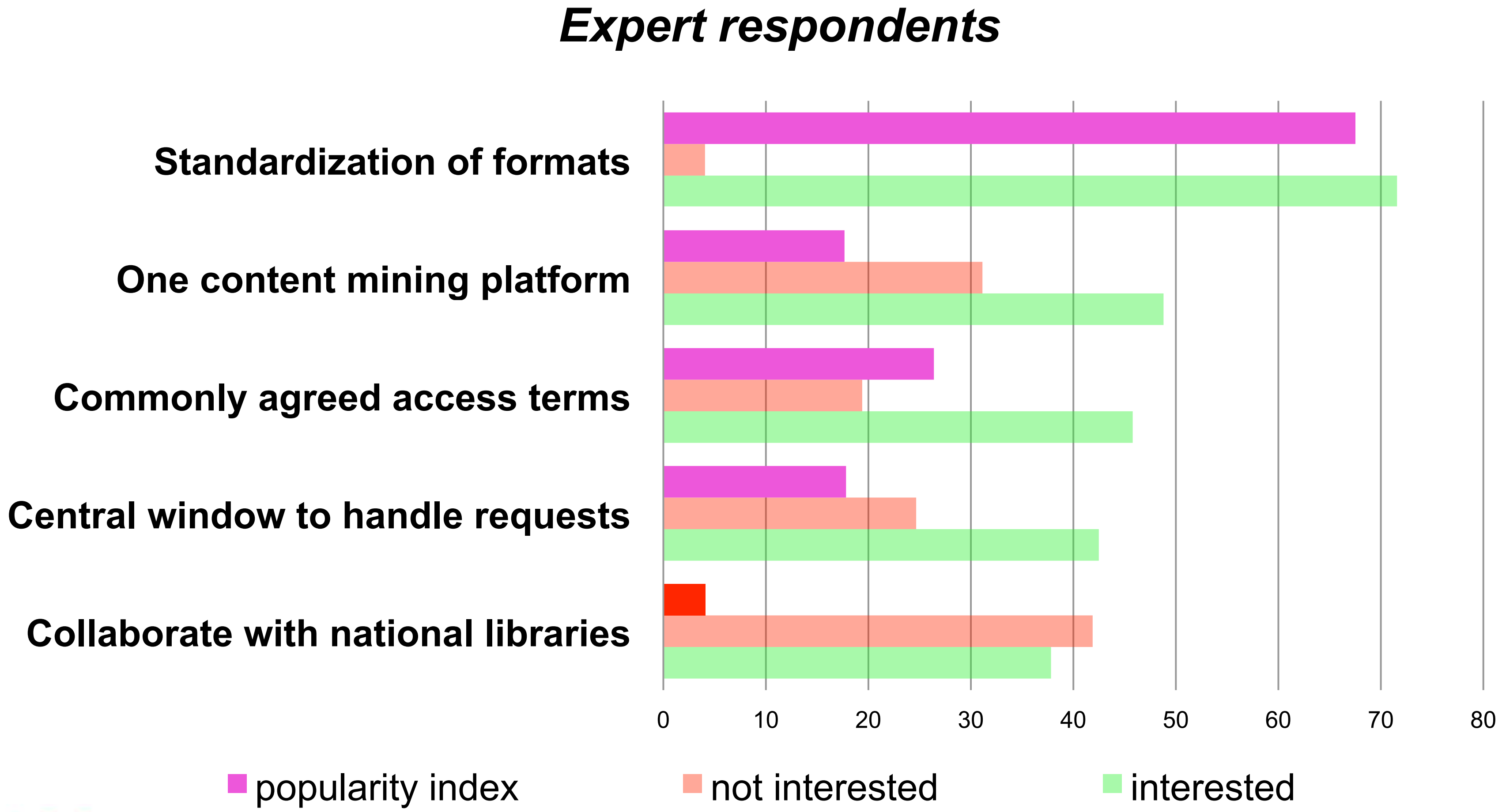
(ad 3: most interviewed experts do NOT see Open Access as a related issue;
access terms also relate to datafile delivery or mining on the platform itself)

Survey results for the 5 suggestions for cross-sector solutions

All respondents



Survey results for the 5 suggestions for cross-sector solutions: Experts



Standardisation best prefered, of content formats (and of API’s)

Experts believe less in one platform and support standardisation even stronger, not just for content, also for APIs:

Top 3 for all Respondents:

1. Standardisation of Formats
2. One Mining Platform
3. Agreed Permission Terms

Top 3 for Experts only:

1. Standardisation of Formats
2. Agreed Permission Terms
3. One Mining Platform

Questions ?

Eefke Smit

Director Standards and Technology

International Association of STM Publishers

smit@stm-assoc.org

stm