



## IMPROVING UPTAKE OF TEXT AND DATA MINING IN THE EU

### Facts

Project No: 665940

Program: H2020 | CSA | GARRI

Duration: 09/2015 - 08/2017

## PaperHive

A start-up perspective



### What does PaperHive do?

PaperHive is a web-platform for collaborative reading that was created in 2016 by Dr. André Gaul and Alexander Naydenov. It allows researchers to engage in collaborative reading which makes reading more effective and efficient. Researchers can easily discover, share and annotate content from different content providers. PaperHive is part of the startup incubator of the Centre for Entrepreneurship at TU Berlin.

### What are the benefits for researchers?

#### 1. More productive reading.

By having a platform that enables research to be understood more easily through comments and discussions, researchers are able to devote more time to other academic literature of interest or other research activities.

#### 2. More learning opportunities.

There are multiple applications in university lectures

and seminars. As people document their questions, thoughts, and ideas within a text, future readers have the opportunity to learn from these documented insights. The value and impact of articles is increased.

#### 3. Increased visibility for both authors and readers.

The interactions taking place through collaborative reading are also opportunities for sharing and networking. In addition to authors gaining more exposure when their articles are commented on, readers that respond to an article have the opportunity to share and raise awareness about their own work when relevant.

### What challenges does PaperHive face as a TDM start up?

#### Content and licensing

Because the current legal regulatory framework around TDM is unclear, PaperHive is unable to roll out additional functionality for the majority of the content, such as



“Publishers should clearly indicate what can be done with the content and not create individual and home-brewed licenses.”

*Dr. André Gaul, CEO PaperHive*

- ▶ TDM services on full-text instead of only abstracts
- ▶ Improving search results by searching in the full text instead of only the metadata
- ▶ Adding additional services such as recommendations based on TDM

Another challenge is the use of different licenses or the absence of clear communications about what can be done with the content across publishers and repositories. Often licenses state the right to ‘read’ the content but not whether this includes TDM or what can be done with the results. Contracts with publishers may allow mining but are unclear about what can be done with the results of the TDM project.

### Technical and infrastructure

**Data formats:** Because PDF is the most widespread format in use, Paperhive uses this format for the platform, even though this is problematic for text and data mining.

**Data quality:** PaperHive does not include articles and books with incomplete metadata because it impairs the user experience.

### What practices for TDM does PaperHive recommend?

**Data quality:** the quality of the existing metadata should be improved significantly by including:

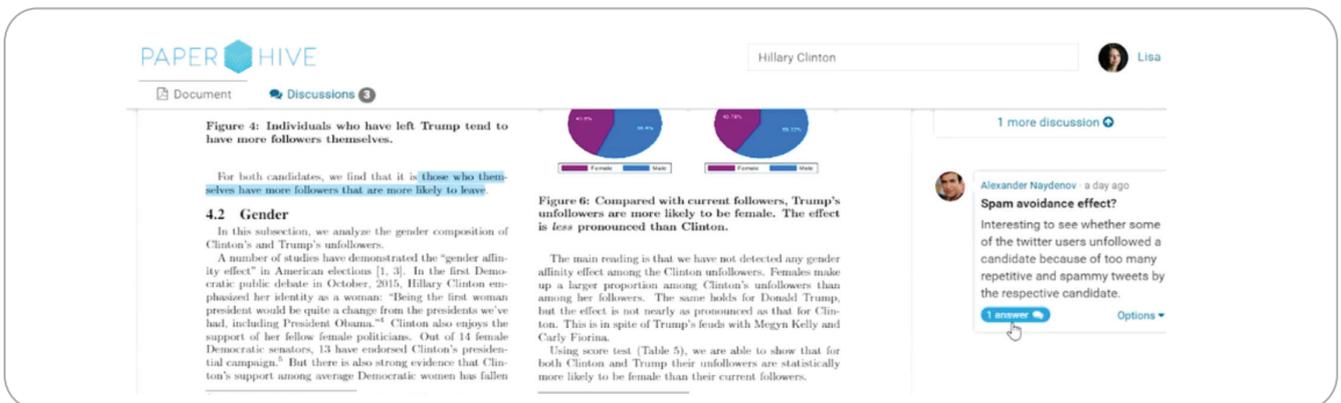
- ▶ Relevant basic information such as author or title
- ▶ Links to full texts in the metadata of all articles and books
- ▶ Information about the format of the full text (e.g. PDF, EPUB, HTML)

**Licensing:** TDM start-ups need licensing information that states clearly what can be done with the data.

▶ Publishers should ideally clearly indicate what can be done with the content and not create their own rules. It is impossible to respect all these different rules on one platform, as this would mean having to deal 7000 different rules and licenses.

▶ Using the CC-BY license for content is a good development. The CC licenses are clear and people in general know what they stand for.

▶ It would be helpful if there was one kind of open data license being adopted as a standard by the community, without additional exceptions.



The screenshot shows a document page on the PaperHive website. The document title is "Hillary Clinton" and the author is "Lisa". The document content includes a figure caption: "Figure 4: Individuals who have left Trump tend to have more followers themselves." and a section titled "4.2 Gender" which discusses the gender composition of Clinton's and Trump's unfollowers. There are two pie charts showing gender distribution. A discussion section is visible on the right, with a comment from Alexander Naydenov titled "Spam avoidance effect?".

*PaperHive website*

Discover more  
VISIT OUR COLLECTION

STORIES 

PROJECTS 

ORGANISATIONS 

TOOLS 

STUDIES 

