Natural Language Processing in Elsevier:

Topic Pages Story

Data Science in Elsevier

Using new capabilities (machine learning, natural language processing, AI) to increase our content utility

Data Science

- What we do

Turn Unstructured Content into Structured Content

- Text Mining
- Images
- Video

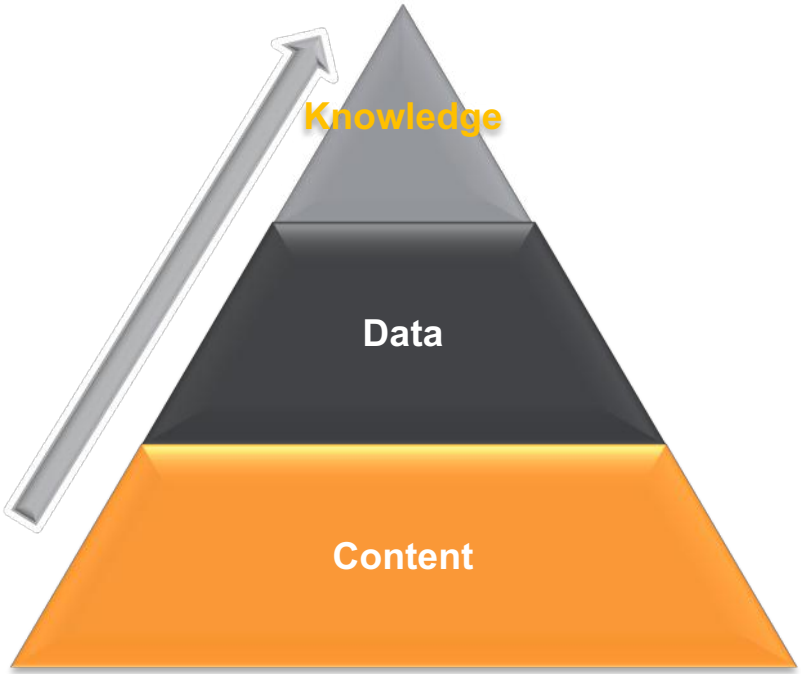
→ Enabling Data Mining
→ Enabling Data Analytics

- Who does it?
  - The team and skills

ELSEVIER

Data Science in Elsevier

Enrichments for evolving information needs and delivery



Answers: *users wanting knowledge – tailor cut to the exact needs of the moment. next-generation search and recommendation Evolved expectations by emergence of AI, Knowledge Graph, new UXes*

Data: *accumulated, structured knowledge. Meta-data around the known entities (authors, articles, geographicals, references, institutions, concepts, relations) – human or machine generated*

Content: *the underpinning of anything good – published material from Journals, Patents, Web, client data.*

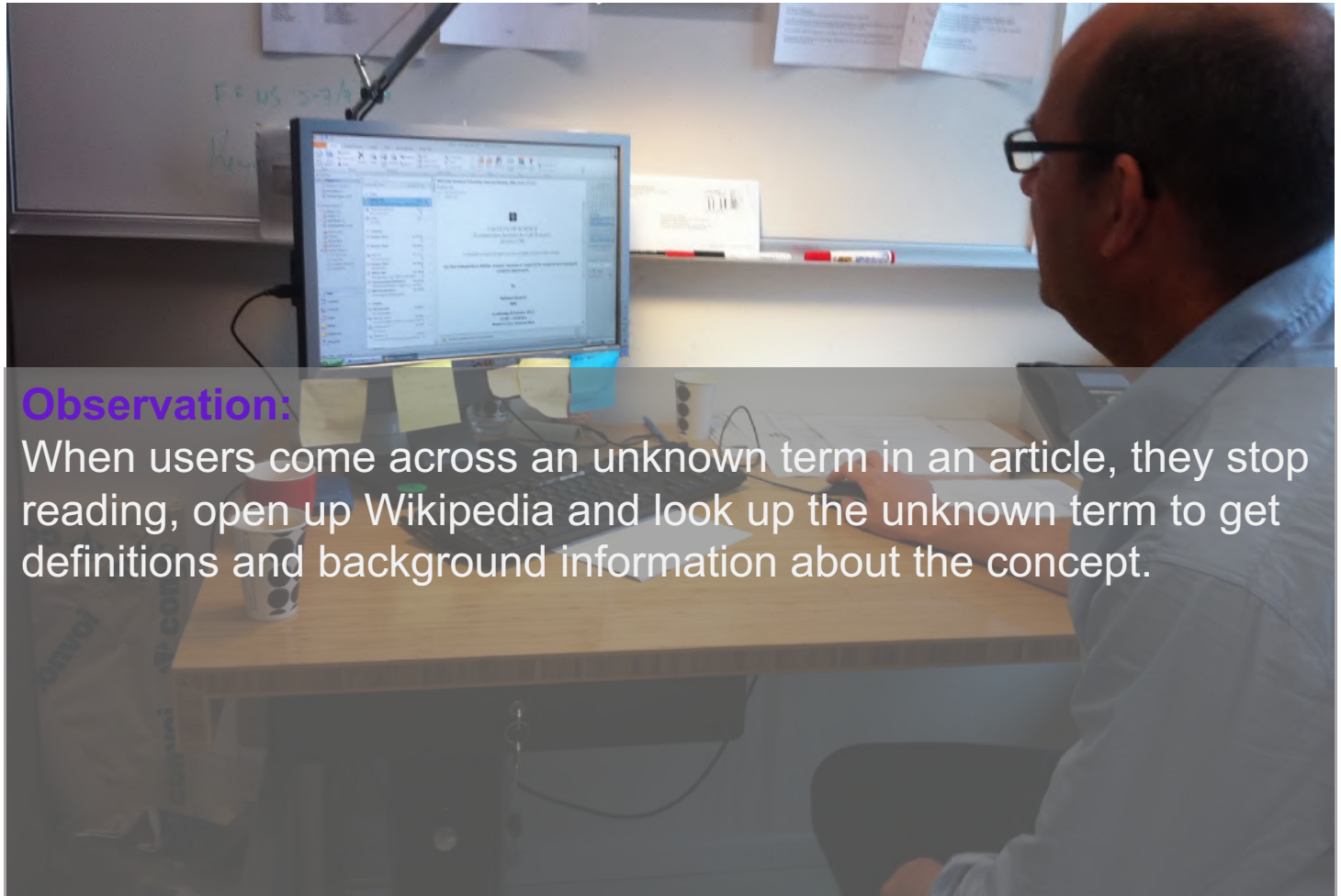
We focus on the enrichments of our content in the production workflows

Science Direct Topic Pages

ELSEVIER

Empowering Knowledge

Key Use Case: Understand the Article



Observation:

When users come across an unknown term in an article, they stop reading, open up Wikipedia and look up the unknown term to get definitions and background information about the concept.

ELSEVIER

ScienceDirect Topic Pages: Case Study

Problem

- Academic articles have scientific concepts
- Researchers need information about unfamiliar concepts they encounter
- They lose time searching for foundational information that is trusted and citable

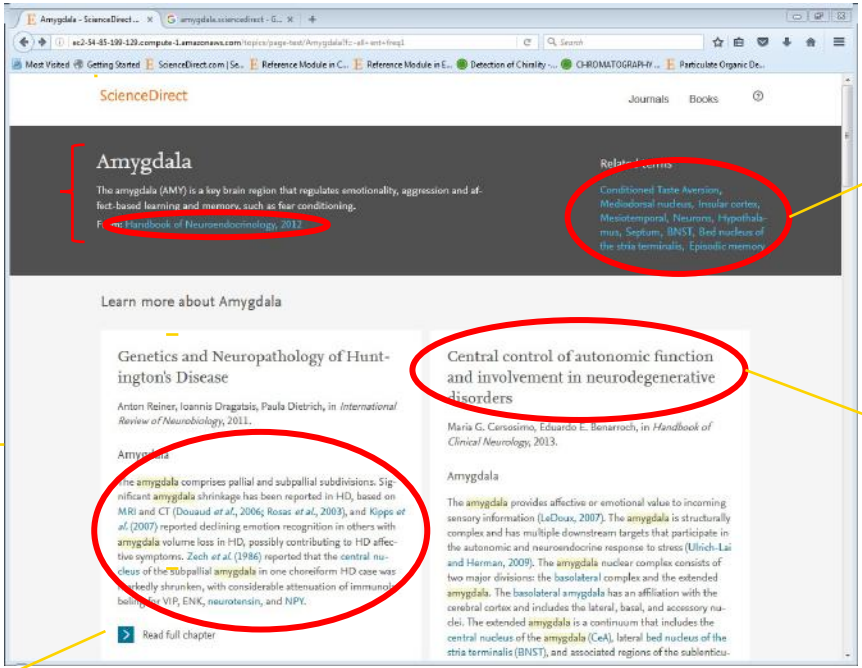
How

- Summarize relevant content from ScienceDirect on *Topic Pages*
- Enrich content with links to the *Topic Pages*
- **Automated** to make processing the content scalable
- Automation presents its own challenges:
  - **Disambiguation** of terms
  - Extraction of **good definitions**

ELSEVIER

Anatomy of a topic page

Definition, clearly delineated



Related terms link to further topic pages *drives serendipity*

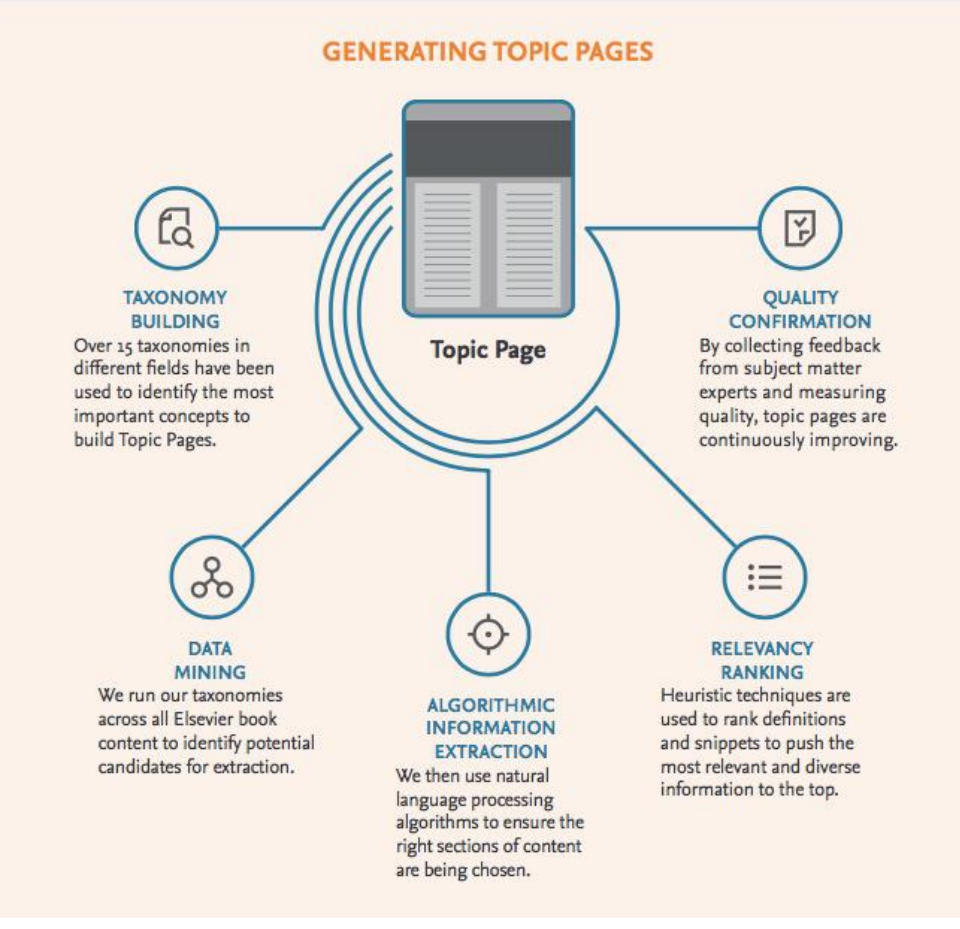
Card presentation supports easy scanning and short snippets preferred by users, *saves time*

Title links to chapter, *drives usage*

”Read full chapter” links at end of snippet, *drives usage*

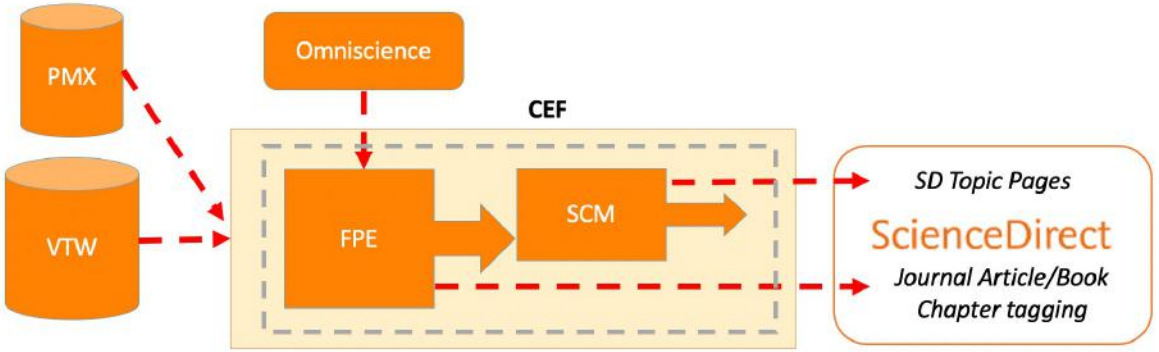
ELSEVIER

Not possible without advanced tech and high quality content



ELSEVIER

Process workflow



Content Enrichment Framework (CEF)

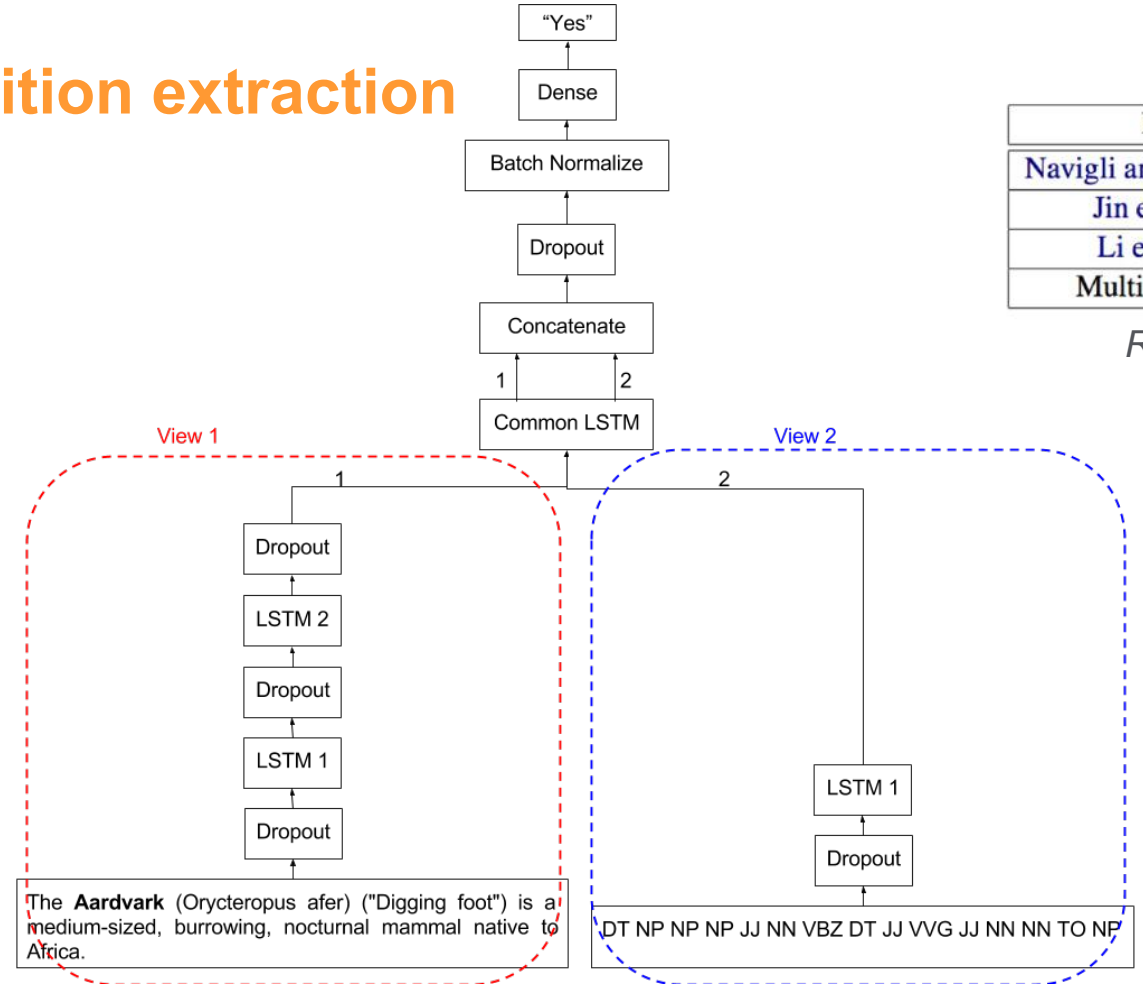
PROCESS WORKFLOW

1. **VTW** – to retrieve the content/XML
2. **PMX** - for metadata and domain classification
3. **Fingerprint engine (FPE)** – Rule based annotation of content using domain specific taxonomies from **Omniscience**
4. **Smart Content Module (SCM)**
   - Definition Algorithm
   - Snippet Algorithm

   Natural language processing and machine learning techniques for relevancy ranking
5. **ScienceDirect (SD)** – Deliver content definitions and snippets

ELSEVIER

Definition extraction



Method	P	R	F1
Navigli and Velardi (2010)	**98.8**	60.7	75.2
Jin et al. (2013)	92.0	79.0	85.0
Li et al. (2016)	90.4	*92.0*	*91.2*
Multi-view LSTM	*95.7*	**95.9**	**95.8**

Results on a public dataset

Learning

- AI models combined with simple annotation tools can uncover the rich knowledge hidden in large unstructured data

- Building high quality topic pages (or any AI-based product) is an iterative process.
  - Build a reasonable product, let users interact, collect feedback, do error analysis and post-hoc interviews, improve and adjust models, and iterate

ELSEVIER

Thank you