



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

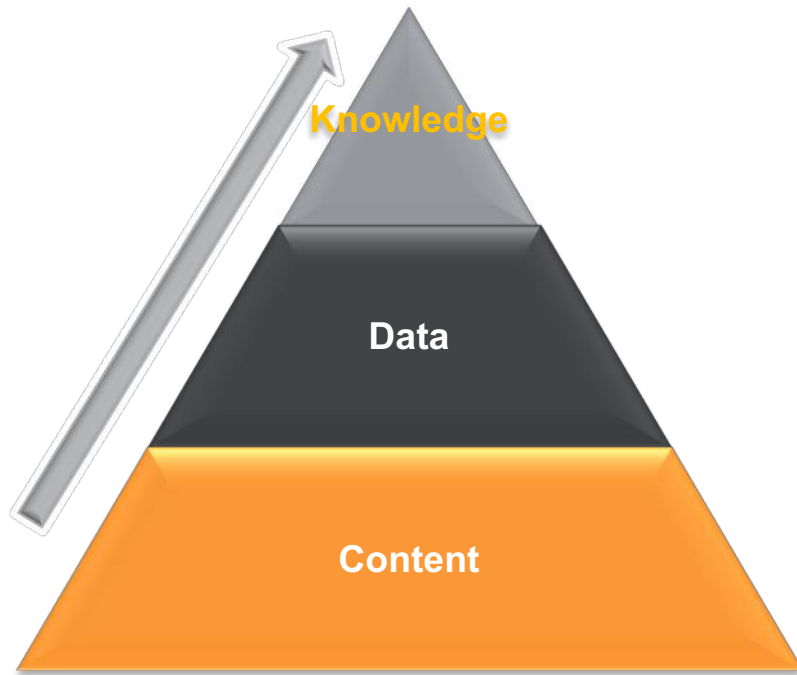
- Who does it?

- The team and skills



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*

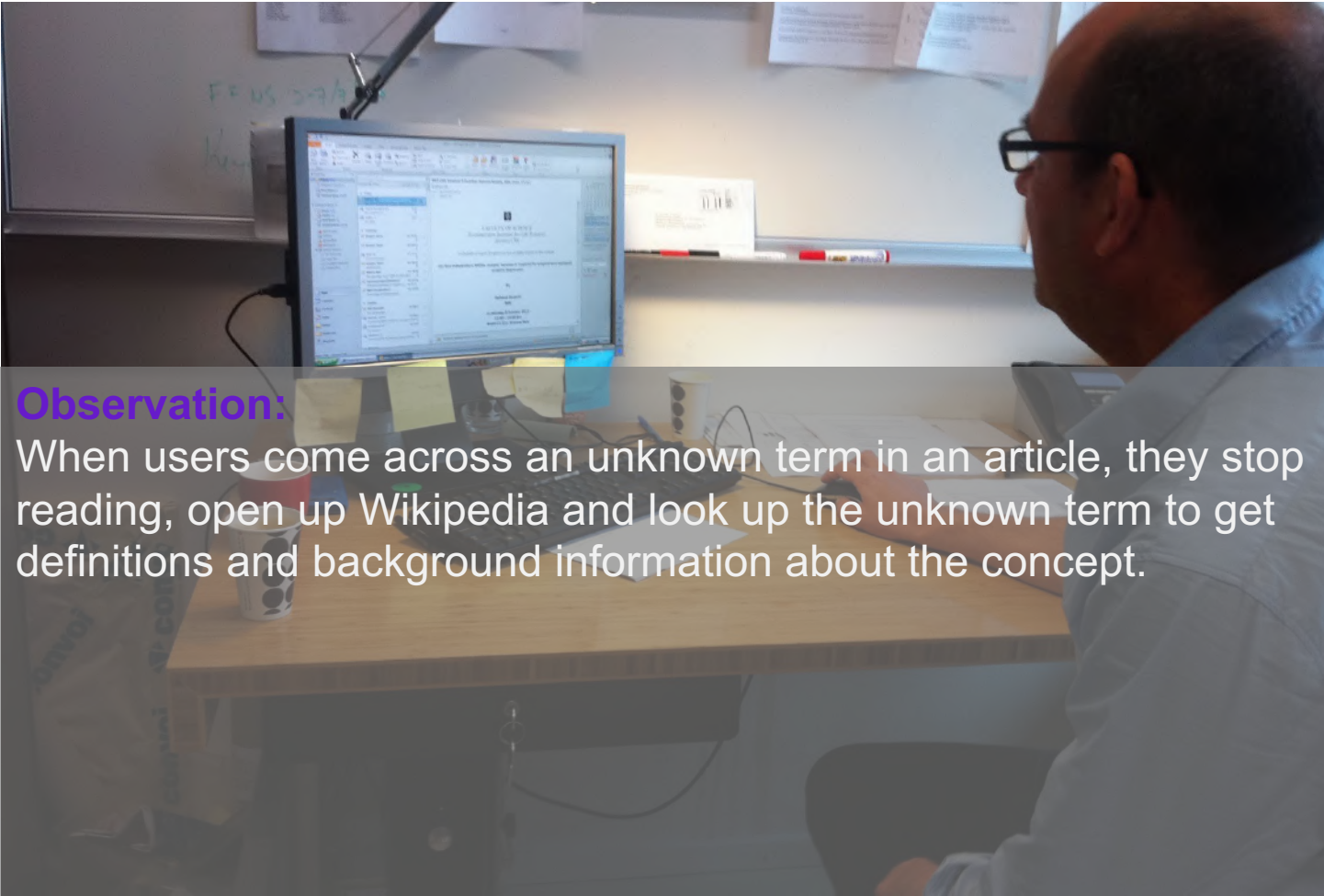
Content: *the underpinning of anything good –
published material from Journals, Patents,
Web, client data.*



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

Science Direct Topic Pages

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.

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**

Anatomy of a topic page

Definition,
clearly
delineated

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

The screenshot shows a ScienceDirect topic page for 'Amygdala'. The page layout includes a header with the ScienceDirect logo and navigation links. The main content area features a large card for 'Amygdala' with a definition and a list of related terms. Below this, there are two smaller cards: 'Genetics and Neuropathology of Huntington's Disease' and 'Central control of autonomic function and involvement in neurodegenerative disorders'. The 'Amygdala' card is circled in red, and the 'Central control of autonomic function and involvement in neurodegenerative disorders' card is also circled in red. A red box highlights the 'Read full chapter' link at the bottom of the 'Amygdala' card. A red box highlights the 'Related terms' section, which lists various brain regions and functions. A red box highlights the title of the 'Central control of autonomic function and involvement in neurodegenerative disorders' card. A red box highlights the 'Read full chapter' link at the bottom of the 'Amygdala' card.

Amygdala
The amygdala (AMY) is a key brain region that regulates emotionality, aggression and affect-based learning and memory, such as fear conditioning.
From: *Handbook of Neuroendocrinology*, 2013.

Related terms
Conditioned Taste Aversion, Mediodorsal nucleus, Insular cortex, Mesiotemporal, Neurons, Hypothalamus, Septum, BNST, Bed nucleus of the stria terminalis, Episodic memory

Learn more about Amygdala

Genetics and Neuropathology of Huntington's Disease
Anton Reiner, Ioannis Dragatsis, Paula Dietrich, in *International Review of Neurobiology*, 2011.

Amygdala
The amygdala comprises pallid and subpallid subdivisions. Significant amygdala shrinkage has been reported in HD, based on MRI and CT (Dowd et al., 2006; Rosas et al., 2003), and Kippes et al. (2007) reported declining emotion recognition in others with amygdala volume loss in HD, possibly contributing to HD affective symptoms. Zach et al. (1986) reported that the central nucleus of the subpallid amygdala in one choreiform HD case was markedly shrunken, with considerable attenuation of immunoreactivity for VIP, ENK, neurotensin, and NPY.

[Read full chapter](#)

Central control of autonomic function and involvement in neurodegenerative disorders
Maria G. Carrasquillo, Eduardo E. Benarroch, in *Handbook of Clinical Neurology*, 2013.

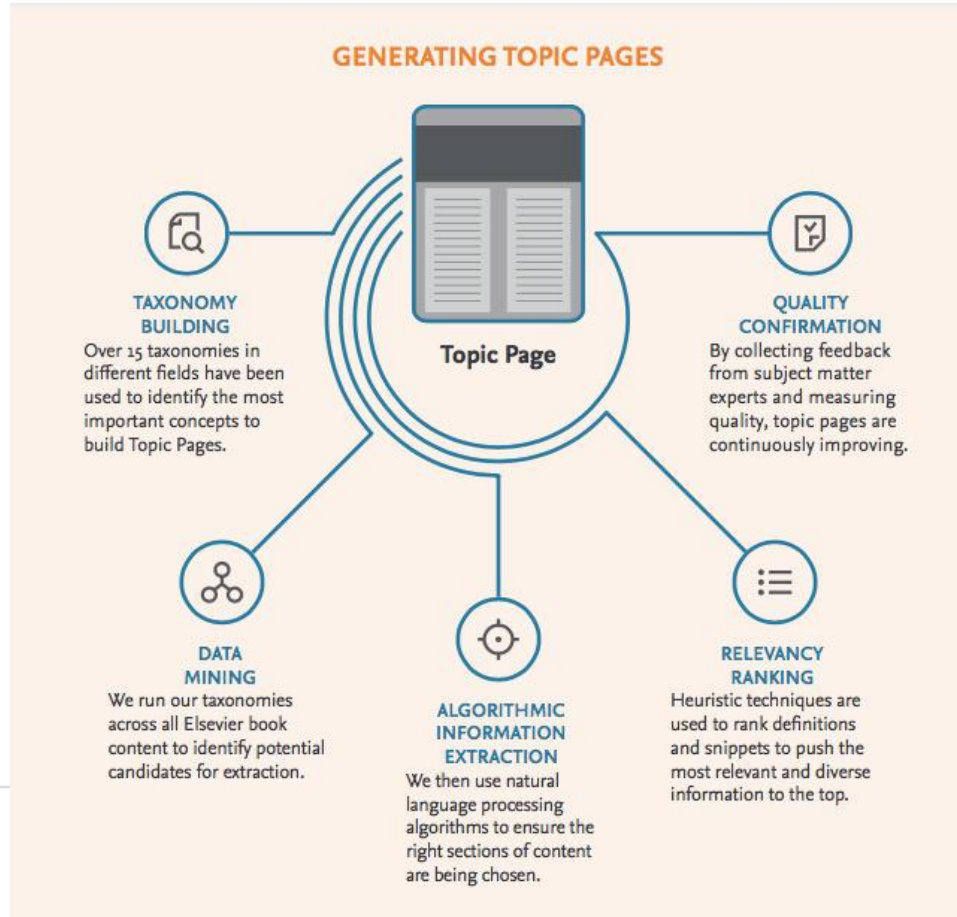
Amygdala
The amygdala provides affective or emotional value to incoming sensory information (LeDoux, 2007). The amygdala is structurally complex and has multiple downstream targets that participate in the autonomic and neuroendocrine response to stress (Ulrich-Lai and Herman, 2009). The amygdala nuclear complex consists of two major divisions: the basolateral complex and the extended amygdala. The basolateral amygdala has an affiliation with the cerebral cortex and includes the lateral, basal, and accessory nuclei. The extended amygdala is a continuum that includes the central nucleus of the amygdala (CeA), lateral bed nucleus of the stria terminalis (BNST), and associated regions of the subnucleus.

Related
terms link to
further topic
pages *drives*
serendipity

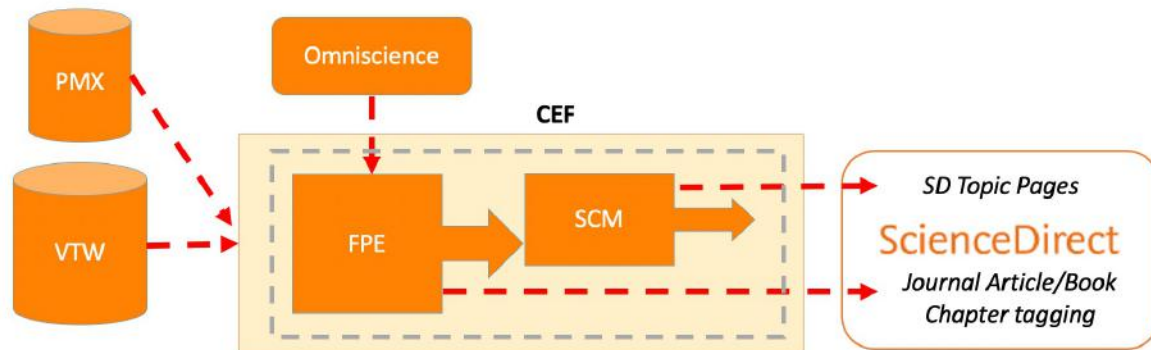
Title links to
chapter, *drives*
usage

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

Not possible without advanced tech and high quality content



Process workflow

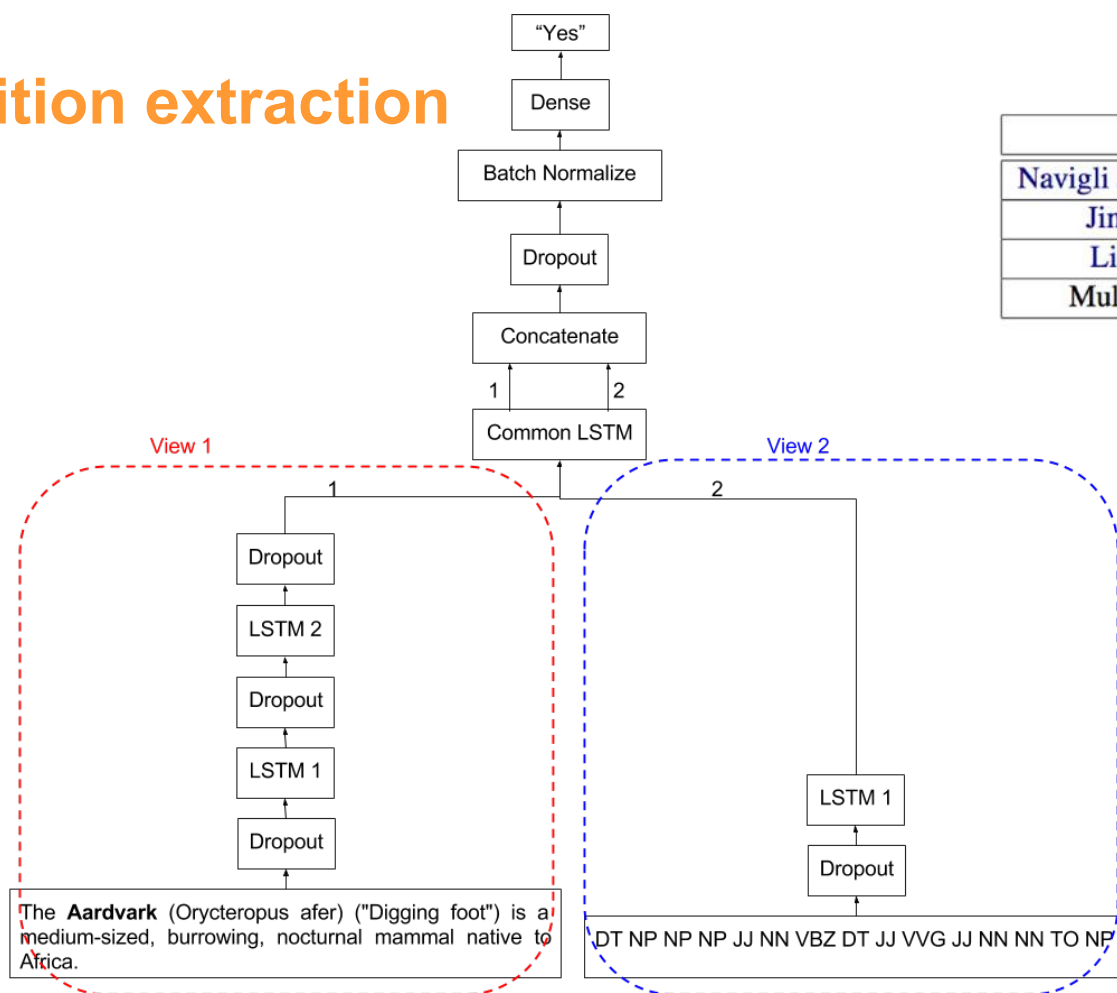


Content Enrichment
Framework (CEF)

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

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

