

Semantics in Scientific Publishing

Approaches & Case Studies

Stefan Geißler, TEMIS Deutschland GmbH

Nordisk konferanse om språk og teknologi

Nasjonalbiblioteket, Oslo, Henrik Ibsens gate 110

7.-8. oktober 2013

Agenda

- Why "Scientific Publishing"? Isn't that hard?
- NLP: Where hard things are sometimes easy and easy things often hard
- Case studies
- TEMIS: Who we are



Publishers of Scientific Journals

Source: 2011 study by the Publishing Research Consortium

Observation

NLP is different from many other technology fields

- NLP (e.g. Semantics) almost mainstream today in scientific publishing
- But in our everyday life today, there's not a lot NLP yet

- Often technology uptake is the other way round
 - First, "simple" requirements can be met
 - Then, as technology advances, sophisticated expert use cases can be handled, too

Examples



First: Simple snapshots only



First: Barely leaving the ground



Today: Complex devices for complex use cases



Today: Broad range of specialized, highperformance machines

NLP is different ...

... because language is different

- Requirements of the "experts" are often easier than modelling presumably everyday communication
- Example I: "Perplexity" of natural language in speech recognition:
 - Speech recognition of radiology reports: Far easier than everyday communication.
 - Counter-intuitive at first: Isn't the special language of highly trained expert more complex?
 - Taking perplexity as the measure, normal newspaper corpora have a perplexity of >100, radiology reports may be as low as 10
 - (Definition: $PP(W) = \sqrt[N]{\frac{1}{P(w_1, w_2, \dots, w_{red})}}$

Example II: Machine translation

- As early as 1981: System METEO: Fully automatic translation of Canadian weather reports, replacing manual translation
- Narrow, expert domain → useful results, clear business case (reduction of efforts, increase in speed)
- But: Machine translation of average personal letters remains challenging until today.

NLP @ TEMIS:
Structure unstructured text
content in business & scientific
contexts to improve how it is
managed and exploited

Key TEMIS customers in Publishing STM, Legal, B2B, Trade, Media





тном son reuters nature publishing group npg





Springer





DE GRUYTER

















































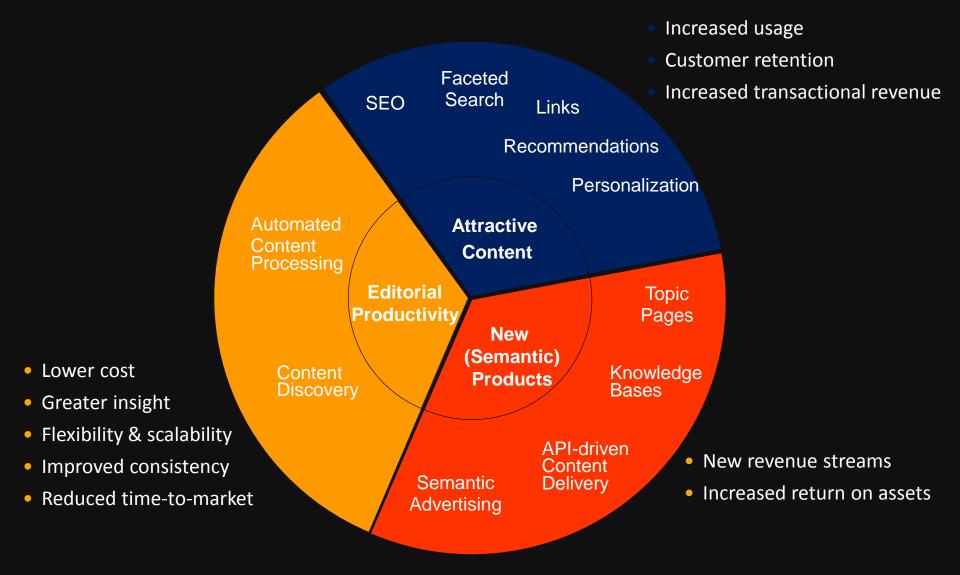




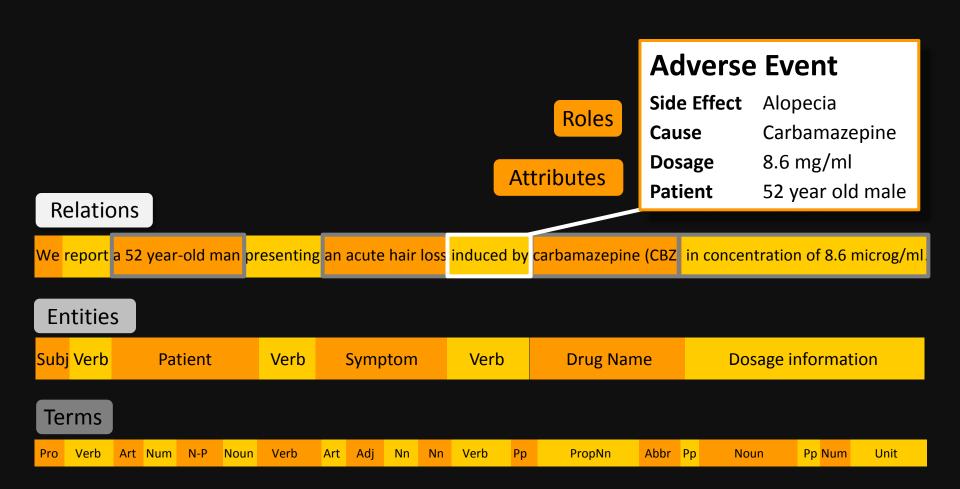




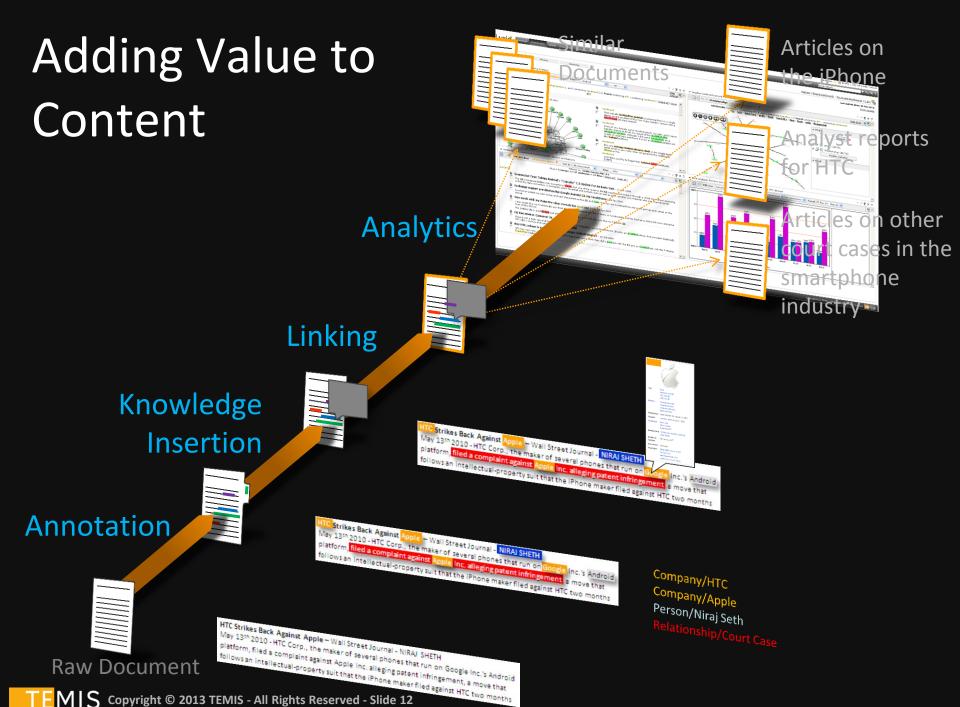
Three Key Areas of Value



What Is Semantic Content Enrichment? The automated extraction of domain metadata



We report a 52 year-old man presenting an acute hair loss induced by carbamazepine (CBZ) in concentration of 8.6 microg/ml.



Radbase project at



- Program goal: implement semantic navigation
 - Giving radiologists access to high quality content enriched with Semantics: Faceted Search, Topic Pages
 - Related Document Linking
- Program motivations
 - Add value to existing content
 - Use Semantics to implement a radiologists workflow (Finding → Diagnosis → Therapy)
- The solution
 - Luxid® Content Enrichment Platform and custom Skill Cartridge® based on Thieme Medical Thesaurus (polyhierarchic & bilingual, 53.000 concepts / 340.000 terms / 350+ MB



Bildschirmmodus:



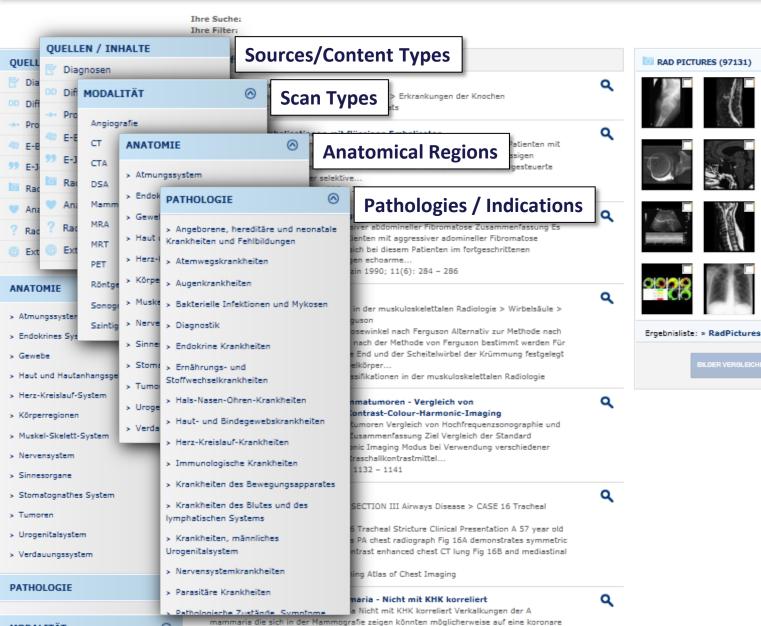
A

Home

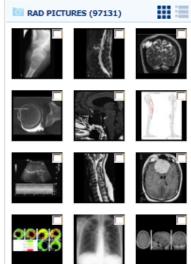
MODALITÄT

My RadBase

Suche



andronikung KHK bindautan Die Studienensbeiere biesen eind allerdiese mideen



Case Law in Numbers project



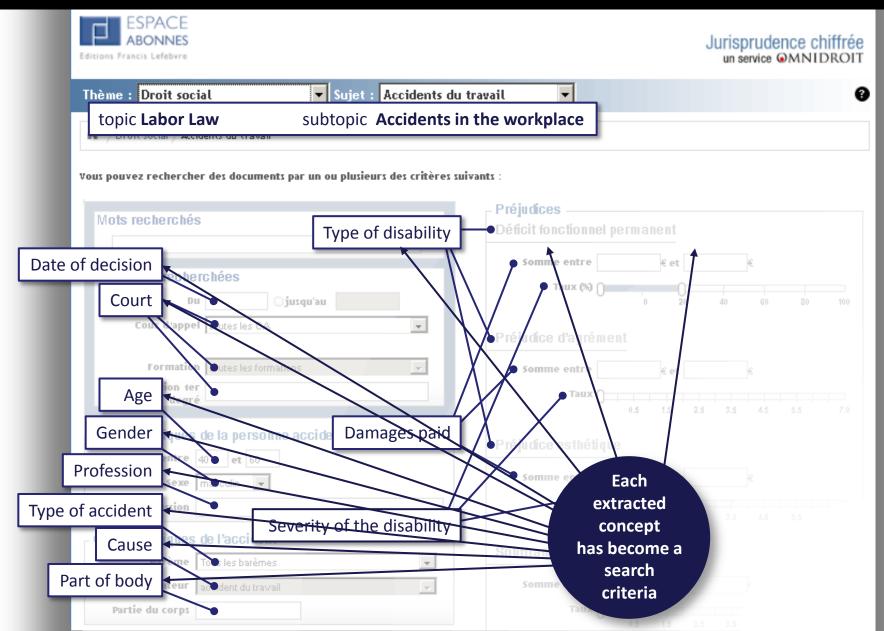
- Program goals
 - Create an interactive knowledge base

by extracting from unstructured legal decisions documents highly relevant information regarding typical amounts in the contexts of: Labor Law (depending on age, company size, location, industry, region, ...) Damages (the type of loss, circumstances, region, ...) Commercial Rent (size, location, facilities, time, region, ...) Alimony/Child support (income, marriage, children, region, ...)

- Program motivations
 - Revenue generation
- The solution
 - Luxid® Content Enrichment Platform
 - Custom-developed Skill Cartridges®
 embedding semantic extraction rules specific to the legal domain

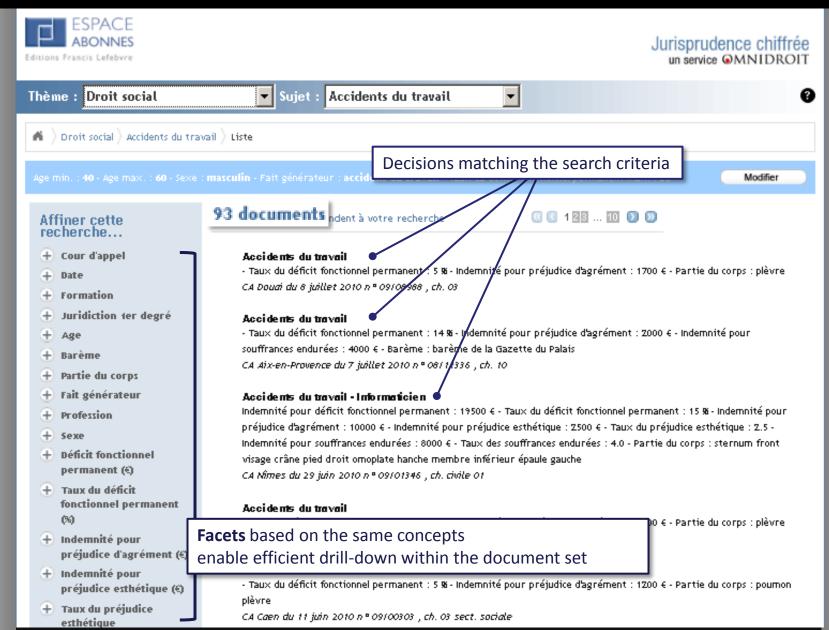
Querying the Knowledge Base





Navigating Query Results





Navigating within 1 query result





Droit social Accidents du travail Liste Consultation d'un document





CA Lyon du 16 septembre 2009 n° 08/00879, ch. 09 IC

Opcument précédent Document suivant

Pointer sur...

- Indemnité pour préjudice d'agrément (€)
- Fait générateur
 accident du travail
 accident de la circulation
- Déficit fonctionnel permanent (€)
- Age
- Taux de souffrances endurées
- Profession
 Technicien
 Pompier
- Sexe
- Indemnité pour souffrances endurées (€)
- Taux du déficit fonctionnel permanent (%)
- Partie du corps

crâne
abdomen
coude gauche
genou droit
fémur
visage
menton
hanche
membre inférieur
cheville
orteil

- Taux du préjudice esthétique
- Indemnité pour préjudice esthétique (£)

Attendu que s'agissant des souffrances endurées, l'expert a relevé que le pretium doloris, proposé par lui au taux de 5/7, tenait compte de la longue évolution, de l'atteinte bilatérale au niveau des membres inférieurs avec des difficultés de déambulation importantes, de l'atteinte faciale avec ses conséquence initiales sur l'alimentation et des conséquences psychiques initiales avec des troubles

initialement importants, ayant nécessité une prise en charge ;

Attendu que M. Raphaël B. justifie aux pièces de son dossier avoir suivi deux séances d'entretien avec un psychologue Mme D., et selon certificats médicaux des 7 mars, 30 avril et 25 mai 2007 du docteur Anne Marie B. son médecin traitant, avoir dû suivre un traitement pour soigner son état dépressif réactionnel;

Attendu qu'au vu de ces éléments, les souffrances endurées paraissent avoir été assez importantes ; que la cour estime, en fonction des justifications produites et en réformant partiellement le jugement de ce chef, devoir indemniser ce poste de préjudice par l'allocation d'une indemnité de 15 000 € ;

Attendu qu'au titre du déficit fonctionnel permanent, l'expert a retenu un taux d'incapacité permanente partielle de 10 % qui tient compte de la raideur modérée de la cheville droite, avec gêne dans les mouvements du gros orteil des douleurs des hanches et des genoux sans limitation fonctionnelle, de l'hypoesthésie des lèvres en rapport pour l'alimentation au niveau des hémi lèvre supérieure gauche et inférieure droite et de la gêne à l'ouverture buccale complète;

Attendu que la victime était âgée de 36 ms au moment de la consolidation (le 11 janvier 2006) ; qu'en lui allouant une indemnite de 11 230 € au titre de ce préjudice, le tribunal a fait une juste appréciation des faits de la cause et qu'il convient de la confirmer ;

Attendo que la Caiss aucune demande de caisse n'établisse pa accident du travail q assuré en tout ou pa

Enhanced experience with accelerated navigation

- Semantic shortcuts accelerate navigation
- Key concepts highlighting

Attendu que le préjudice d'agrément est constitué d'après l'expert, par le fait que les séquelles de l'accident, conservées par la victime, entraîneront une plus grande pénibilité de la marche prolongée et des difficultés pour pratiquer le VTT de façon intensive, toutefois sans contre indication à la pratique en dilettante de cette activité;

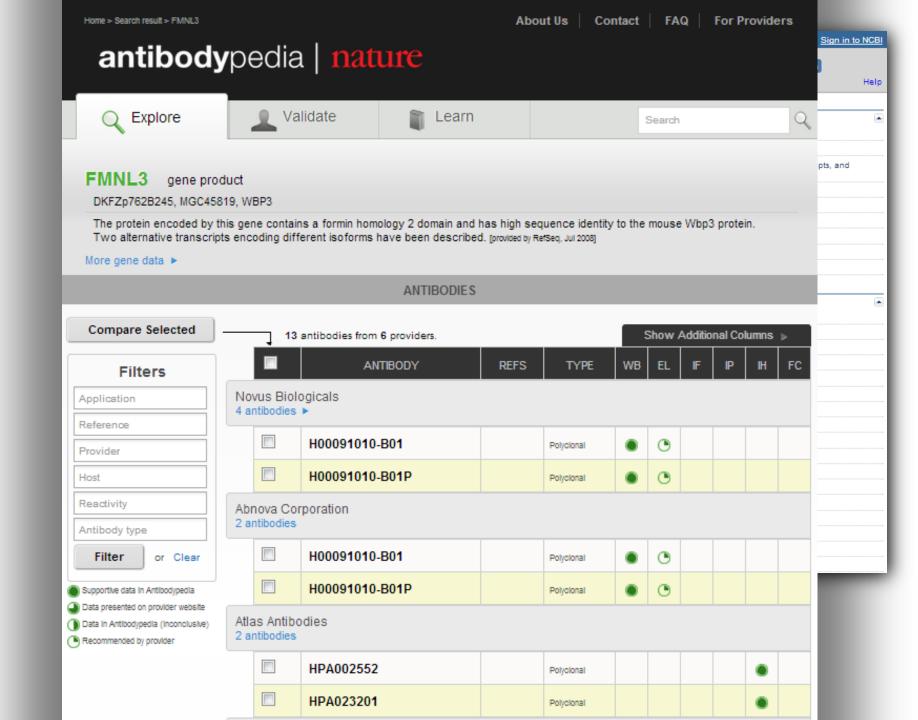
Attendu qu'en l'état de ces éléments, et sans qu'il soit besoin d'exiger la production de plus amples justificatifs eu égard à la pature du sport pratiqué, en tenant compte également de l'abandon forcé de l'activité de pompier.



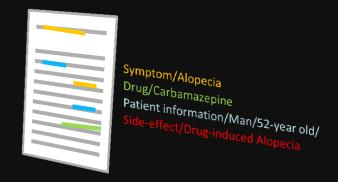
Smart Content at



- Program goals
 - Make content more compelling
 - Provision new and evolving products and services
 - Make publishing more dynamic
- Program motivations
 - Customer retention
 - Revenue generation
- The solution
 - Luxid® Content Enrichment Platform
 - Life Sciences Skill Cartridges®: Chemistry, Medicine and Biology

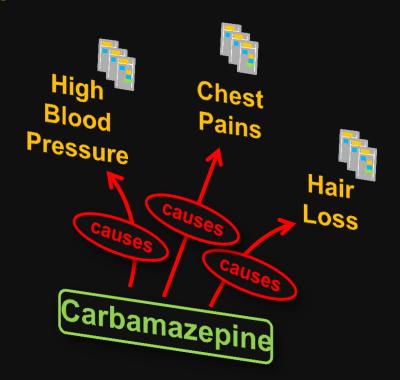


Towards data-centric In addition to Content-centric



Content-Centric

"which documents could you show me that discuss side effects of carbamazepine?"



Data-Centric

"what are the side effects of carbamazepine?"

Workflow Productivity Example metrics



Automated indexing of 2M abstracts

5 months vs. 2,000 man months (manual)



Volume of indexed documents (stable indexing team size)



Automated linking of legal citations Anonymization of case law



Resources required for legal consolidation Daily vs weekly update of the law

Luxid: The TEMIS Approach to Semantics

Platform approach

Annotation Factory

Skill Cartridge[®] Library

Content Enrichment Studio

Generic Analysis Engine: Multilingual Part of Speech Tagging, Scalability, REST Web Service API, UIMA, Rules Engine, Document Categorization, Clustering, ...

Collection of generic or domain- / client specific analysis components implementing vocabularies, rules, statistical models to extract specific types of information from text

Integrated development environment to develop, test, evaluate, maintain and extend analysis components, syntax-aware, one-click compilation, deployment, collaboration, ...







TM360° People Companies Locations



BER **Biological Entities** Relationships



Events Military **Political** Strategic



Competitive Intelligence Business

Relationships



MER Medical **Entities** Relationships



Opinion Mining **Opinions** Sentiment



IPTC News Categories



CER Chemical **Entities** Recognition



RTF STF

•••

One Skill Cartridge Example: Terminology Matching

<skos:broader rdf:resource="#2132" />



Thesaurus of Agricultural terms, used at the US National

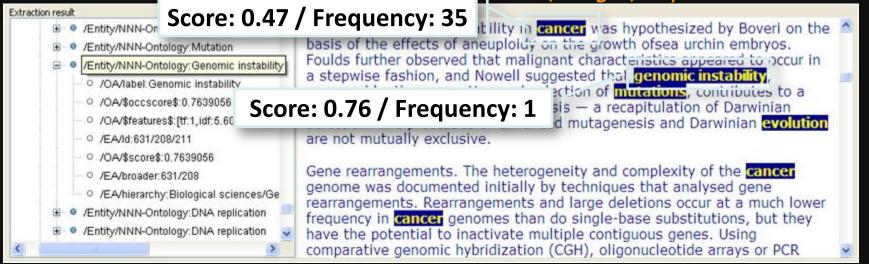
```
<skos:Concept rdf:about="#29126">
  <skos:prefLabel xml:lang="en">corn</skos:prefLabel>
  <skos:prefLabel xml:lang="es">maíz</skos:prefLabel>
  <skos:scopeNote xml:lang="en">Use for corn as a grain; For corn as a vegetable USE sweetcorn; For the corn plant USE Zea
    mays.</skos:scopeNote>
  <skos:scopeNote xml:lang="es">Usar para maíz en grano; USAR "maíz dextrinoso" para referirse al maíz como hortaliza; USAR "Zea mays"
   para designar la planta.</skos:scopeNote>
  <skos:altLabel xml:lang="en">feed corn</skos:altLabel>
  <skos:altLabel xml:lang="en">field corn</skos:altLabel>
  <skos:altLabel xml:lang="en">maize</skos:altLabel>
  <skos:altLabel xml:lang="es">maiz</skos:altLabel>
                                                                  Related term
  <skos:related rdf:resource="#29135" />
  <skos:related rdf:resource="#6414" />
ckos:Concent rdf:shout="#10919"
 <skos:prefLabel xml:lang="en">grains</skos:prefLabel>
<skos:Concept rdf:about="#6414">
 <skos:prefLabel xml:lang="en">nixtamalization</skos:prefLabel>
 <skos:prefLabel xml:lang="es">nixtamalización</skos:prefLabel>
 <skos:definition xml:lang="en">Nixtamalization involves cooking and steeping corn in a lime solution, then washing and grinding the corn
   to form masa, a dough used to make tortillas, taco shells, corn chips, etc.</skos:definition>
 <skos:definition xml:lang="es">La nixtamalización consiste en cocer y remojar maíz en una solución de cal, luego el maíz se lava y muele
   para formar una masa con la que se hacen tortillas, hojarazcas para tacos, chips de maíz, etc.</skos:definition>
 <skos:altLabel xml:lang="en">nixtamalisation</skos:altLabel>
 <skos:altLabel xml:lang="en">nixtamalising</skos:altLabel>
 <skos:altLabel xml:lang="en">nixtamalizing</skos:altLabel>
 <skos:hiddenLabel xml:lang="en">nixtamalised</skos:hiddenLabel>
 <skos:hiddenLabel xml:lang="en">nixtamalized</skos:hiddenLabel>
 <skos:altLabel xml:lang="es">nixtamalizacion</skos:altLabel>
 <skos:related rdf:resource="#6413" />
 <skos:related rdf:resource="#29126" />
 <skos:related rdf:resource="#29149" />
 <skos:related rdf:resource="#29173" />
```

One Skill Cartridge Example: Terminology Matching



- Indexing
 - Build a specific Skill Cartridge® for thesaurus indexing, based on the NAL thesaurus
 - Increased <u>precision</u> by computing the relevance of terms

based on position, frequencies (document, reference corpus), measure of connectedness with other terms, length, depth ...



One Skill Cartridge Example: Terminology Matching



- Indexing
 - Build a STF-based (Smart Taxonomy Facilitator) Skill Cartridge®
 - Increased <u>precision</u> by computing the relevance of terms

based on position, frequencies (document, reference corpus), measure of connectedness with other terms, length, depth ...

Increased <u>recall</u> by applying fuzzy term matching techniques
 Identifies and scores lexical variations of thesaurus terms

	Term	Score
Preferred form	Acousto-optical deflector	1.0000
Permutation	Optic and acoustic deflector	0.8677
Insertions	Acousto-optic based deflector	0.8339
Spelling alternatives	Acousto-optical deflecter	0.9259

Skill Cartridges Flexible approach to address many use cases

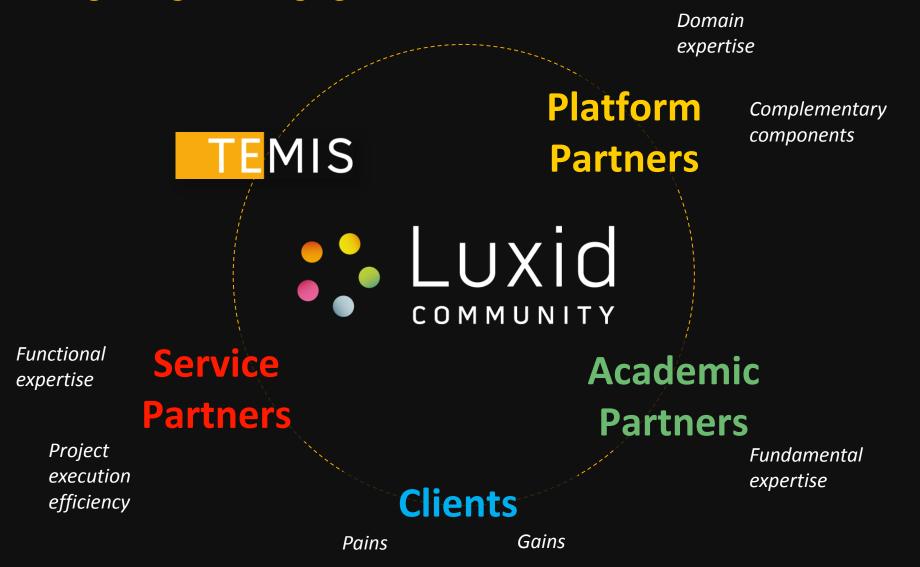
- Generic software platform Luxid + specific Skill Cartridge
 - Allows addressing many different use cases indexing scientific documents in life sciences, legal, engineering, ... sentiment analysis, patent analysis, ...
 - There are more opportunities than TEMIS can address alone
- Invite 3rd parties to participate
 - Define new, additional Skill Cartridges
 - Open license or commercial?
- A marketplace to suggest a technical and formal framework: The "Luxid Community"
 - An "AppStore" for Semantic components
 - Development tools, teaching material, access to existing install base, discuss, learn, exchange



community.temis.com

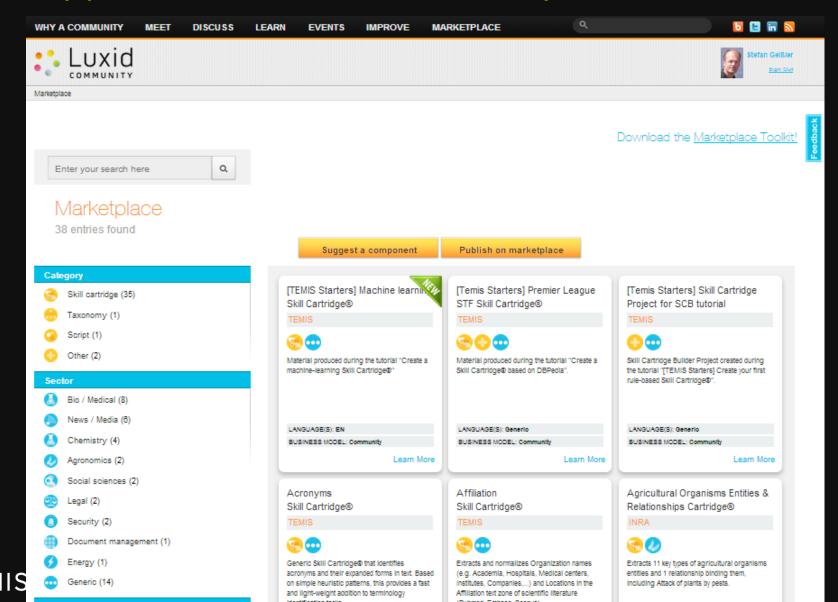
Luxid® Community

Building, using, exchanging resources for Semantic Content Enrichment



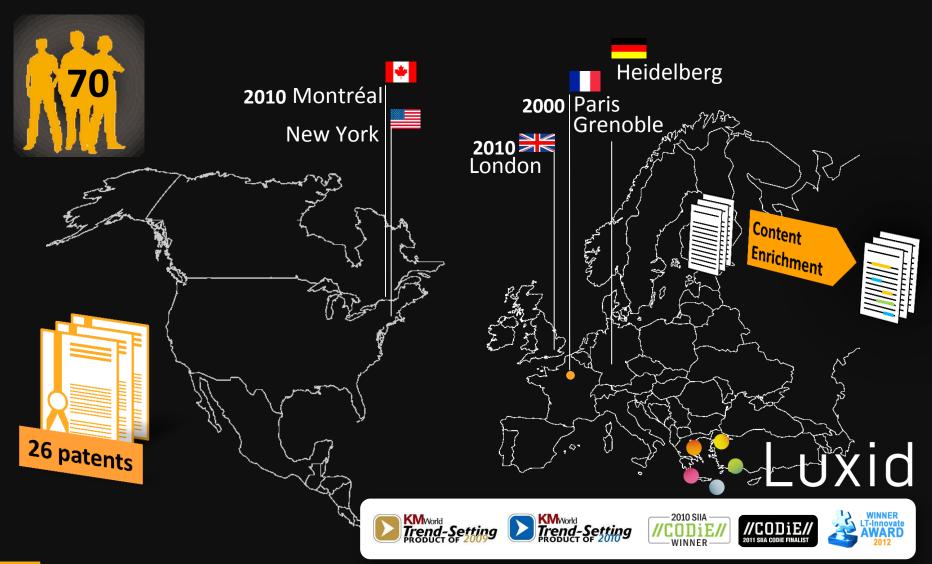
The Luxid Community

An "App Store" for Semantic components



TEMIS

Semantic Content Enrichment software vendor since 2000



Closing thoughts

- Semantics in Scientific Publishing
 - Faster & simpler process in internal editorial workflows
 - Additional insight & perspective beyond the raw content
 - New types of products, possible only with automated semantics
- TEMIS approach
 - Luxid Platform: One software core, many components
 - Multilinguality, scalability, APIs, rule-based, thesaurus-based, machine-learning, ...
- Luxid Community: Meeting the interests of different stakeholders
 - Luxid user: More existing, ready-made domain-specific components
 - Domain expert ("App Developer"): Access to industry clients
 - TEMIS: Increased reach of the Luxid approach



Takk for oppmerksomheten!

TEMIS Deutschland GmbH Blumenstraße 15 69115 Heidelberg

www.temis.com

Stefan Geißler

TEMIS Chief Innovation Officer
+49 (62 21) 1 37 53 - 12
+49 (174) 6 59 57 12

stefan.geissler@temis.com