Semantic Research Service

for unstructured documents

Alexandre Quemy, Eren Dabanlioglu, Piotr Walczak

June 13, 2016

IBM Analytics

PLAN

The Idea and Business Value

Market Landscape and Size

Business Model

The Architecture

The demo

THE IDEA AND BUSINESS VALUE

THE INITIAL PROBLEM

Initial observation

Many complex real life decisions require:

1. Access to knowledge from different sources.
2. Collect, connect, and exploit knowledge .

The problems

1. Too much data to be processed efficiently by a human.
2. Keywords search engines are limited.

$\Longrightarrow$ research time is prohibitive,
knowledge is very incomplete,
global analysis is partial.

The product idea

Smart search engine that understands the **underlying** structures of documents:

- reduce research time.
- provide more relevant information than keywords by semantic similarities coupled with *features* and hints extraction.
- provide specific Analytics and decision aid tools for a given domain.
- Natural Language Processing capabilities.

$\Longrightarrow$ better decisions, reduction of risks.

THE IDEA AND BUSINESS VALUE

Not only a search engine:

- **Corpus complement**: what are the most relevant documents to match a collection of documents?
- **Corpus explanation**: how my documents are connected?
- **Forecasting models** for crutial questions related to the field of my corpus*.
- Usage of **public and private** documents:
  - **Public data**: fiscal reports, patents, social network, newspapers, court opinions, etc.
  - **Private data**: customer records, internal emails, etc.
- **Different data types**: text, images, sounds, videos.

* Requires expertise to create ad-hoc models.

USE CASE IN LAW

Risk management

A lawyer has a client A that wonders if a hostile takeover toward B could be perceived as breaking antitrust regulation due to an increasing monopoly and lead to sanctions.

Entity relations

A lawyer has some documents on a case and wants to have as much information on persons involved (relations, CV, public profiles, etc.) as possible. Idem for companies.

USE CASE IN HEALTHCARE

Diagnosis system

A doctor provides the full medical record of a patient and recent MRI results in order to find patients' records with similar medical path and provide more accurate differential diagnosis.

EXTENTION TO OTHER FIELDS

- Finance & Investment
- Engineering
- Journalism
- ...

Each field requires expert knowledge but the **core service** is the same.

Law as first domain

- what is the distribution of compensations for such a case, depending on those criteria?
- what documents, hints can increase my chance to win out?
- how a new law, jurisprudence is being applied?
- ...

MARKET LANDSCAPE AND SIZE

MARKET LANDSCAPE

Potential customers:

Anyone that would need semantic research:

- Law: law firms, law schools, corporations.
- Healthcare: hospital, insurance companies.
- Investment: banks, any investor.
- ... more as we offer a flexible service.

MARKET LANDSCAPE FOR LEGAL ANALYTICS

In US:

1. 1,300,000 licensed attorneys in the United States.
2. 58 million consumers in the U.S. sought an attorney.
3. 200 law schools.

In France:

- 60000 lawyers, **+41% in 10 years**,
- in 2014, **791.448 basic missions** for juridical help.
- 8355 judges
- around 50 law universities
- **legal analytic is a priority axis of development**

MARKET SIZE: A BRIEF REVIEW

”Westlaw and LexisNexis share a market that is reportedly worth **$8 billion** a year, constantly growing, and needing more and more analytics.”

As February 2016:
”The total addressable market for legal software – both corporate law departments and law firms—is **15.9 billion annually**; the market spends $3 billion each year; law departments spend $1.5 billion annually on 11 types of software—from matter management to compliance to legal analytics – in a market with a **$6.5 billion potential** and; while all technology segments are growing.” — InsideCounsel

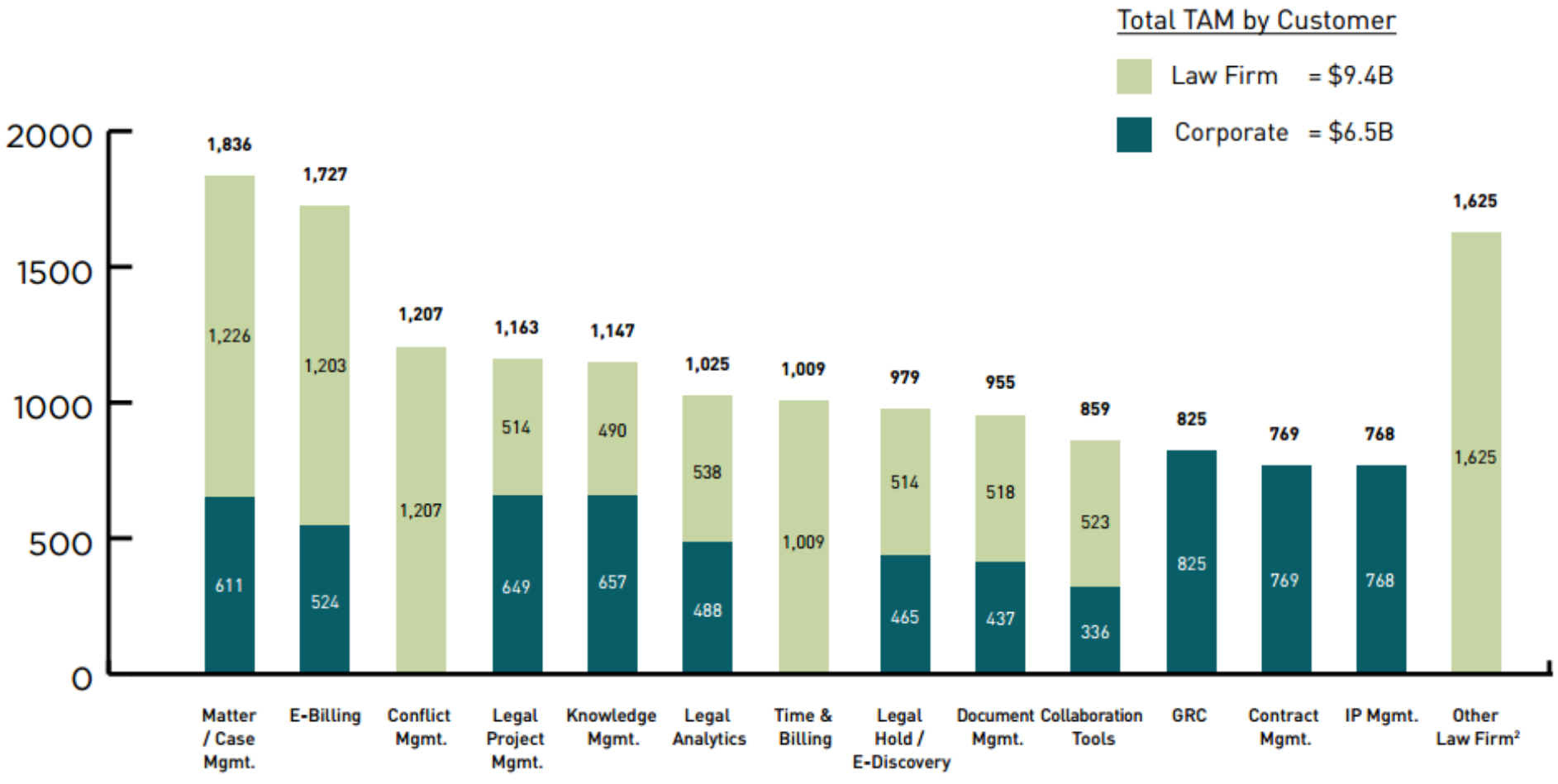
MARKET SIZE: CORPORATE LEGAL SOFTWARE

Corporate Legal Software	2015	2019	CAGR	Total Addressable Market (with 100% penetration)	2015 Penetration
e-Billing	$202m	$235m	4%	$524m	39%
Matter management	$195m	$279m	9%	$611m	32%
Contracts management	$187m	$346m	17%	$769m	24%
Governance & compliance	$147m	$270m	16%	$825m	18%
IP management	$140m	$194m	8%	$768m	18%
Legal hold	$129m	$158m	5%	$465m	28%
Document management	$127m	$183m	10%	$437m	29%
Legal project management	$102m	$198m	18%	$649m	16%
Knowledge management	$99m	$259m	27%	$657m	15%
Legal analytics	$73m	$145m	19%	$488m	15%
Collaboration tools	$61m	$92m	11%	$366m	18%

MARKET SIZE AND SEGMENTS

Full Target Addressable Market (TAM)[1] by Product Type

$ (millions)



Note: [1]Assumes all potential accounts are penetrated (100%) and product spend is the higher end of the range;
[2]Includes calendering/docketing ($574m), legal financial management ($556m), and legal process automation ($23m)

MARKET LANDSCAPE

Competitors:

- In general, no competitor: unique service.
- Could be perceived as competitor by some, but...
- ... can work in synergy (e.g. plugin LexisNexis).

MARKET LANDSCAPE

Strength:

- Licensed data: journals, books,...
- Really fast to add new decisions.

Weaknesses:

- Non-flexible business model.
- Slow and not ergonomic.
- Poor analytics compared to IBM capabilites.

BUSINESS MODEL

FREEMIUM MODEL

Free for basic research:

- keywords only, no corpus completion.
- no account or alert on a research.
- no analytics on documents.
- no private sources of documents.
- limitation on request numbers? restriction to registered users?

Then initial subscription $60 / mo / user for premium features.
+ $XX / mo / user for specific analytic module (one package = one use case)
+ $XX / Go of private sources indexed and analysed.

Pros:

- Users will interact and provide useful data.
- Easier to advertise more advanced products (analytic modules, eDiscovery, ROSS,...).

Cons:

- Slower ROI at first?

STRATEGY

Two-step strategy (per domain):

1. Create the 'free' version, i.e. focus on purely semantic search engine features:
   - 1.1 Index a lot of public data.
   - 1.2 Provide a very fast and efficient search engine.
   - 1.3 Allow the user to feed our database with annotations.

   $\Longrightarrow$ Retain customers and advertise the product
2. Progressively add analytic capabilities, forecast module, premium features, etc.

ETA for Beta: 3 months.

ACADEMIC PARTNERSHIP

Poznan University of Technology

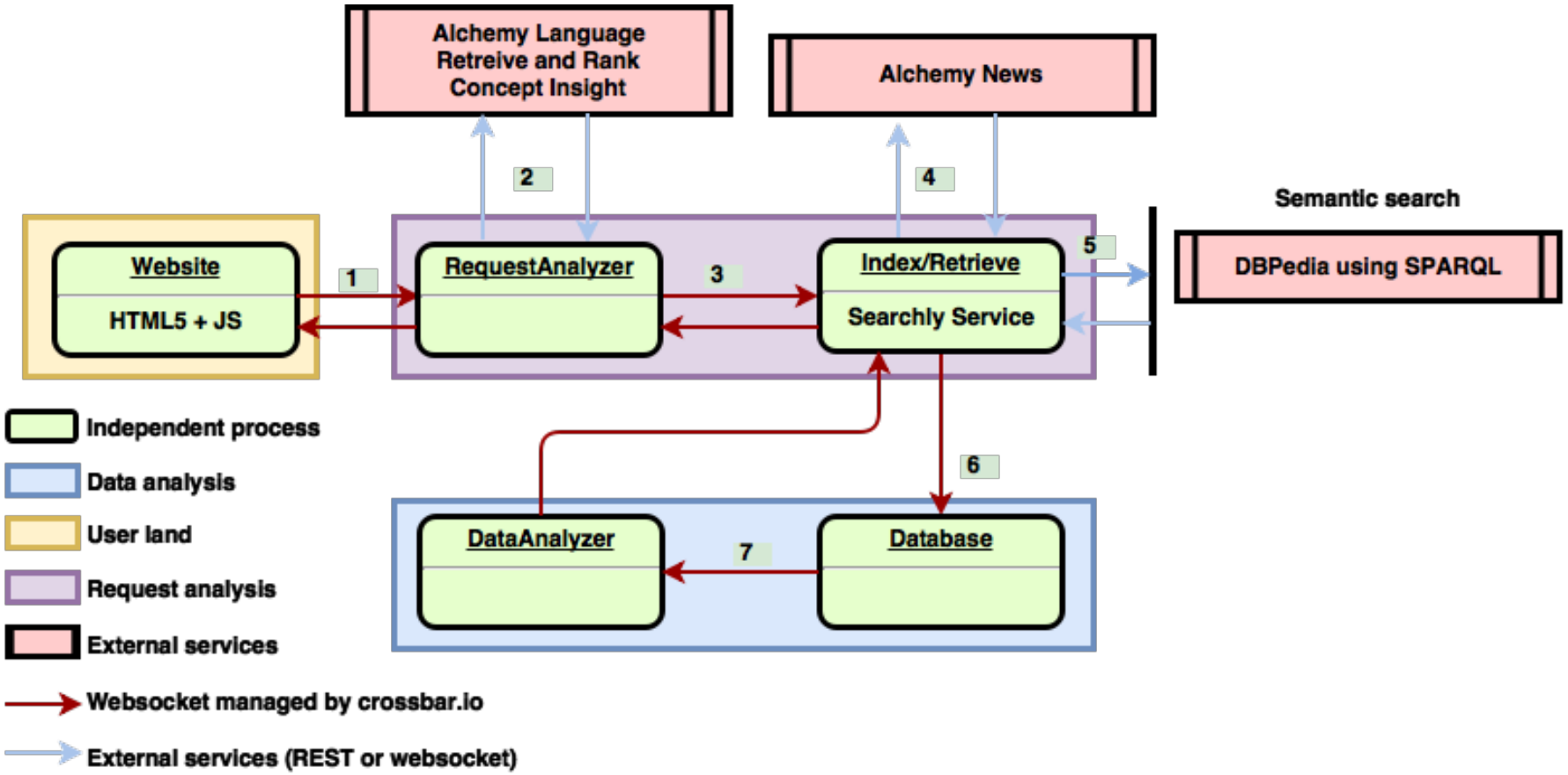
Partnership with the Faculty of Computing:

1. Students could perform Master Thesis for IBM.
2. IBM could outsource projects to PUT.
3. PhD project on Legal Analytics starting.

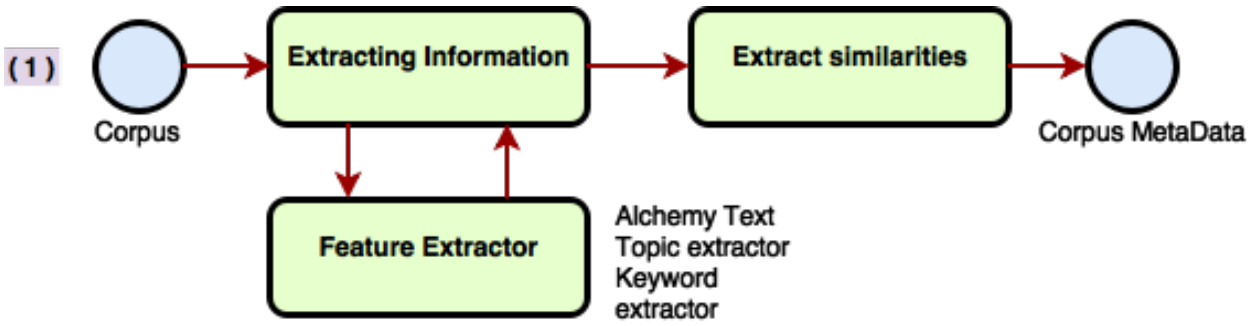
Similar efforts are done in Healthcare domain.

THE ARCHITECTURE

THE BIG PICTURE



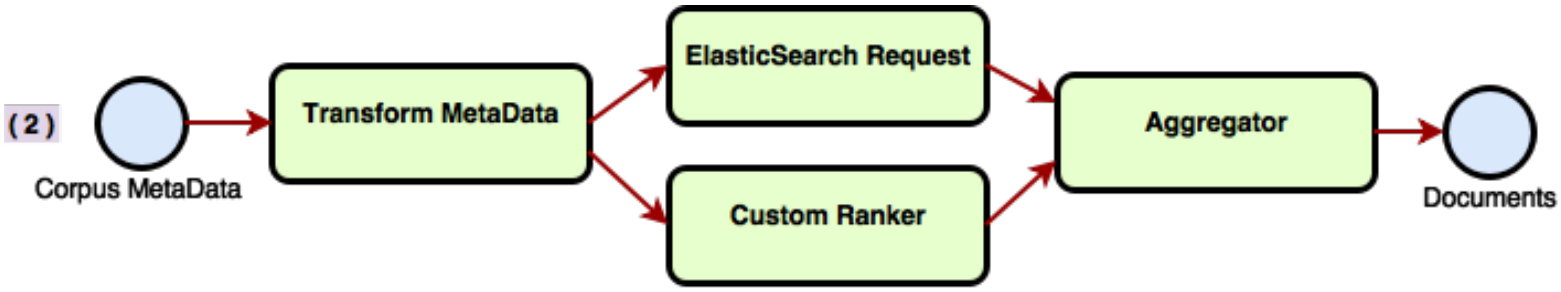
(1) CORPUS ANALYSIS



Implementation:

- Feature extractor: Alchemy Language, Timestamp.
- Corpus of... one document only.
- Draft of similarities computation.

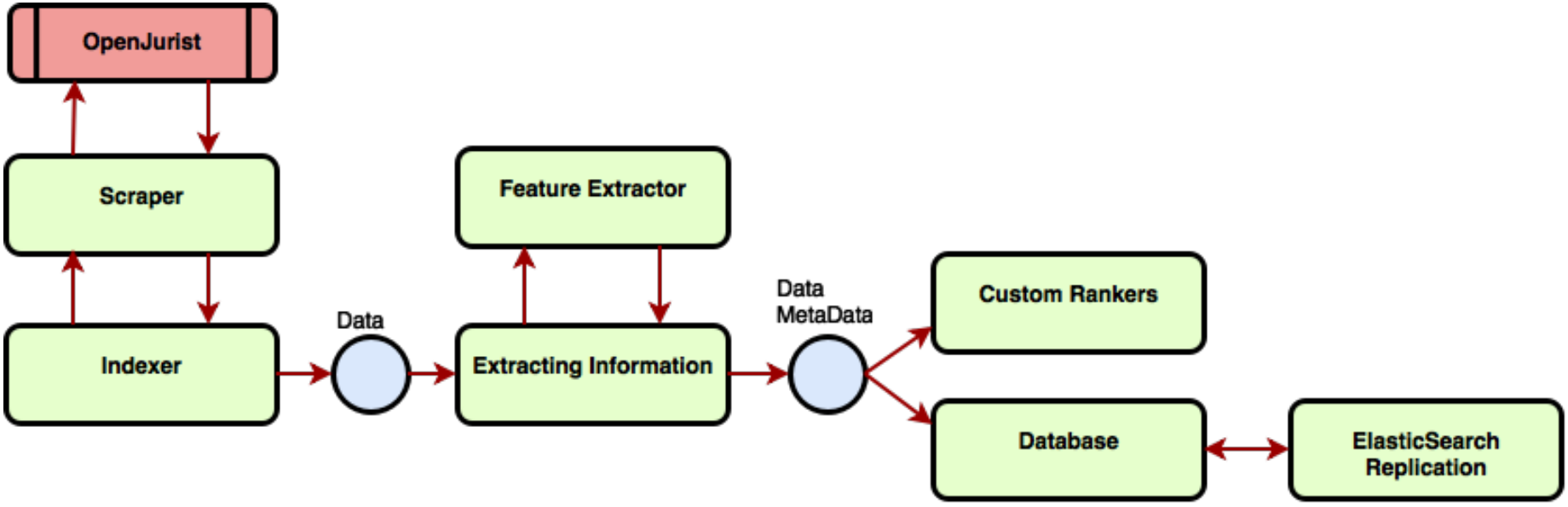
(2) SEARCH AND RANK



Implementation:

- No aggregator (requires more expertise).
- ElasticSearch: Fuzzy request using Alchemy Keywords.
- Custom Rank: Latent Semantic Index.

FUNCTIONAL PROCESS: DOCUMENT INDEXING



Some figures:

- About 7h to collect 30k raw documents from OpenJurist.
- About 14h to extract metadata using Alchemy.
- About 1h to create the models (persistent).
- About 15min to build the matrix (every restart).

THE DEMO

THE END

THE END

Thank you for your attention!
Questions?