Elsevier's Challenge

Dynamic Knowledge Stores and Machine Translation

Presented By Michelle Gregory, Pascal Coupet

Date 17-05-2016

OUTLINE

- **Introduction**
  - Elsevier: from publisher to a data & analytics company
- **Elsevier Data**
- **Elsevier Products**
- **Challenges**
- **Current status on Challenges:**
  - Knowledge Graphs
  - Machine aided translation
- **Challenge details:**
  - Creating high quality knowledge graphs
  - Linking taxonomies to translation memory to support machine aided translation

FROM PUBLISHER TO DATA & ANALYTICS COMPANY

Over the last
50 years
the majority of Noble Laureates have published with Elsevier

Founded over
130 years ago

Employ over
7,000 employees
in 25 countries

Published over
440,000 articles
in 2015

Received over
1.4 million submissions in 2015

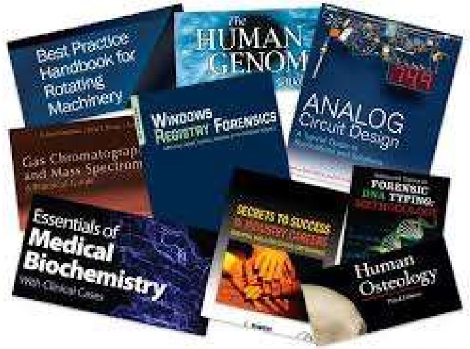
Work with over
30 million
Scientists, students, health & information professionals

Over **61 million**
items indexed by Scopus

	CONTENT	SOLUTIONS			
CAPABILITIES	**Elsevier eBooks, Online Journals, Databases**	**Elsevier Research Intelligence**	**Elsevier R+D Solutions**	**Elsevier Clinical Solutions**	**Elsevier Education**
	Publishes over 2,200 online journals & over 10,000 e-books	Provides universities, governments, and research institutions with the resources and insights to improve institutional research strategy, management, and performance.	Helps corporate researchers, R+D professionals, and engineers improve how they interact with, share, and apply information to solve problems using our digital workflow tools, analytics, and data	Helps medical professionals apply trusted data and sophisticated tools to make better clinical decisions, deliver better care, and produce better healthcare outcomes.	Helps educate highly-skilled, effective healthcare professionals, using the most advanced pedagogical tools and reference works.
	Cell ap THE LANCET Compendex	Pure SciVal	Knovel Geofacets Embase Reaxys	ClinicalKey ToxED	Mosby's Skills+
PLATFORMS	ScienceDirect		Scopus		MENDELEY

ELSEVIER DATA

- **Journals**
  - 3000 journals
  - 440000 articles
  - 1.4 million submissions/year
- **Books**
  - 10000+ eBooks
- **Citations, abstracts and references**
  - 61 million abstracts in Scopus
- **Databases**
  - 26 million substances in Reaxys
  - 4000 drugs in PharmaPendium
  - and more…
- **Taxonomies**
  - 20000 concepts in Omniscience (general subject)
  - 1 million concepts in EMMeT (medicine)
  - 70000 concepts in EmTree (medicine)
  - and more…



Scopus

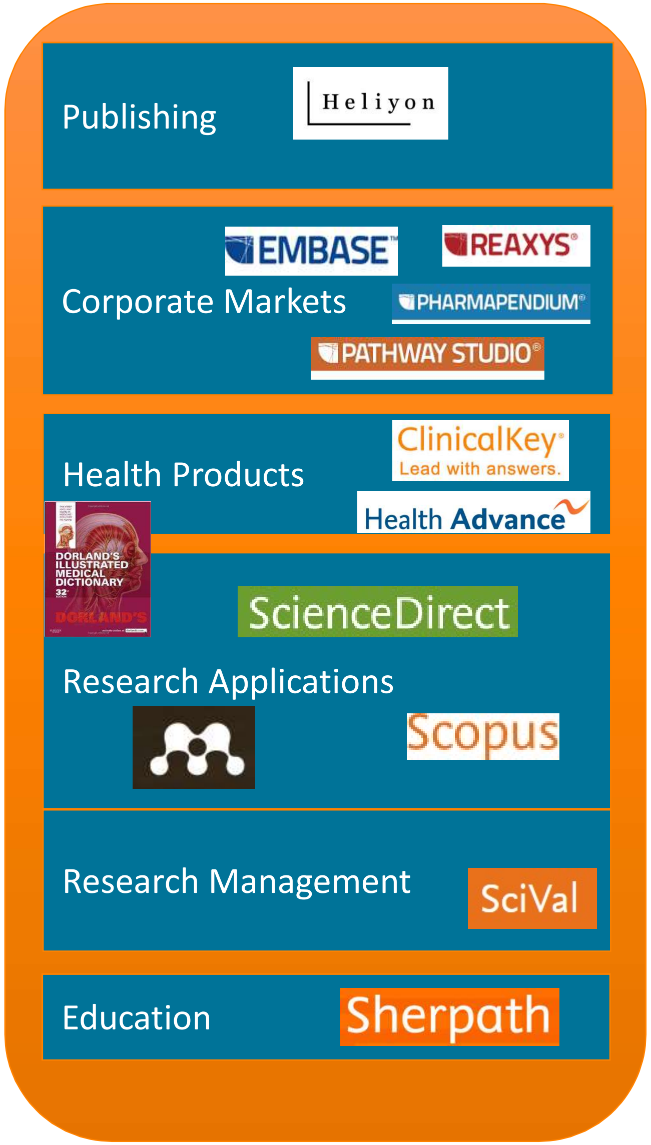
REAXYS®

PHARMAPENDIUM®

OmniScience

ELSEVIER PRODUCTS

- **Platforms:**
  - ScienceDirect
  - Health Advance
  - Mendeley
- **Products based on analytics:**
  - SciVal
  - Pure
- **Products based on curated data:**
  - Reaxys
  - PharmaPendium
  - Engineering Village
  - Geofacets
  - Pathway Studio



Publishing

Corporate Markets

Health Products

Research Applications

Research Management

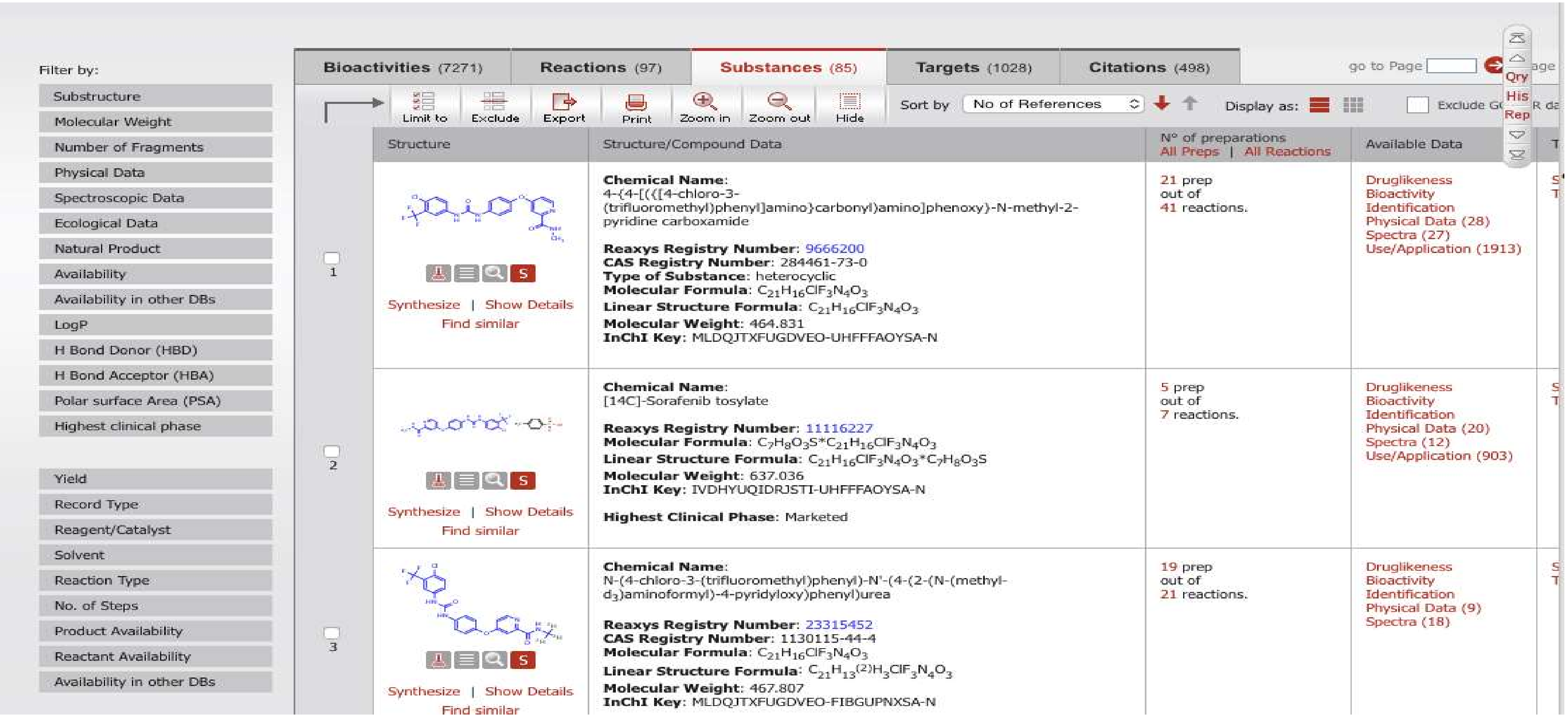
Education

THE CHALLENGES…

1. **How to create high-quality non-trivial Knowledge Graphs?**

2. **Machine Aided Translation:**
   - How to connect/use multi-lingual taxonomies to memory-based translation?
   - How to generate translations of taxonomies?

STRUCTURED DATA – A COMPETITIVE EDGE



Filter by:

- Substructure
- Molecular Weight
- Number of Fragments
- Physical Data
- Spectroscopic Data
- Ecological Data
- Natural Product
- Availability
- Availability in other DBs
- LogP
- H Bond Donor (HBD)
- H Bond Acceptor (HBA)
- Polar surface Area (PSA)
- Highest clinical phase

- Yield
- Record Type
- Reagent/Catalyst
- Solvent
- Reaction Type
- No. of Steps
- Product Availability
- Reactant Availability
- Availability in other DBs

Sort by: No of References

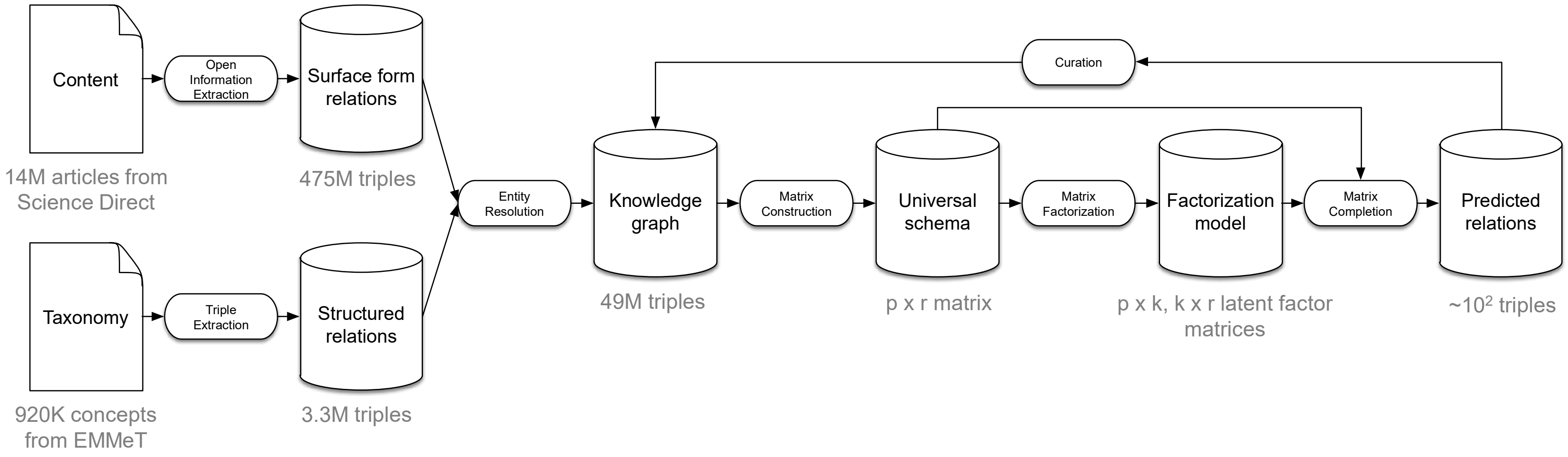
Display as:

□ Exclude G R da

	Structure	Structure/Compound Data	N° of preparations All Preps | All Reactions	Available Data	T
□ 1	Synthesize | Show Details Find similar	**Chemical Name**: 4-{4-[({[4-chloro-3-(trifluoromethyl)phenyl]amino}carbonyl)amino]phenoxy}-N-methyl-2-pyridine carboxamide **Reaxys Registry Number**: 9666200 **CAS Registry Number**: 284461-73-0 **Type of Substance**: heterocyclic **Molecular Formula**: $C_{21}H_{16}ClF_3N_4O_3$ **Linear Structure Formula**: $C_{21}H_{16}ClF_3N_4O_3$ **Molecular Weight**: 464.831 **InChI Key**: MLDQJTXFUGDVEO-UHFFFAOYSA-N	21 prep out of 41 reactions.	Druglikeness Bioactivity Identification Physical Data (28) Spectra (27) Use/Application (1913)	S T
□ 2	Synthesize | Show Details Find similar	**Chemical Name**: [14C]-Sorafenib tosylate **Reaxys Registry Number**: 11116227 **Molecular Formula**: $C_7H_8O_3S*C_{21}H_{16}ClF_3N_4O_3$ **Linear Structure Formula**: $C_{21}H_{16}ClF_3N_4O_3*C_7H_8O_3S$ **Molecular Weight**: 637.036 **InChI Key**: IVDHYUQIDRJSTI-UHFFFAOYSA-N **Highest Clinical Phase**: Marketed	5 prep out of 7 reactions.	Druglikeness Bioactivity Identification Physical Data (20) Spectra (12) Use/Application (903)	S T
□ 3	Synthesize | Show Details Find similar	**Chemical Name**: N-(4-chloro-3-(trifluoromethyl)phenyl)-N'-(4-(2-(N-(methyl-d_3)aminoformyl)-4-pyridyloxy)phenyl)urea **Reaxys Registry Number**: 23315452 **CAS Registry Number**: 1130115-44-4 **Molecular Formula**: $C_{21}H_{16}ClF_3N_4O_3$ **Linear Structure Formula**: $C_{21}H_{13}{}^{(2)}H_3ClF_3N_4O_3$ **Molecular Weight**: 467.807 **InChI Key**: MLDQJTXFUGDVEO-FIBGUPNXSA-N	19 prep out of 21 reactions.	Druglikeness Bioactivity Identification Physical Data (9) Spectra (18)	S T

WHAT WE’VE DONE SO FAR: BUILDING KNOWLEDGE GRAPHS

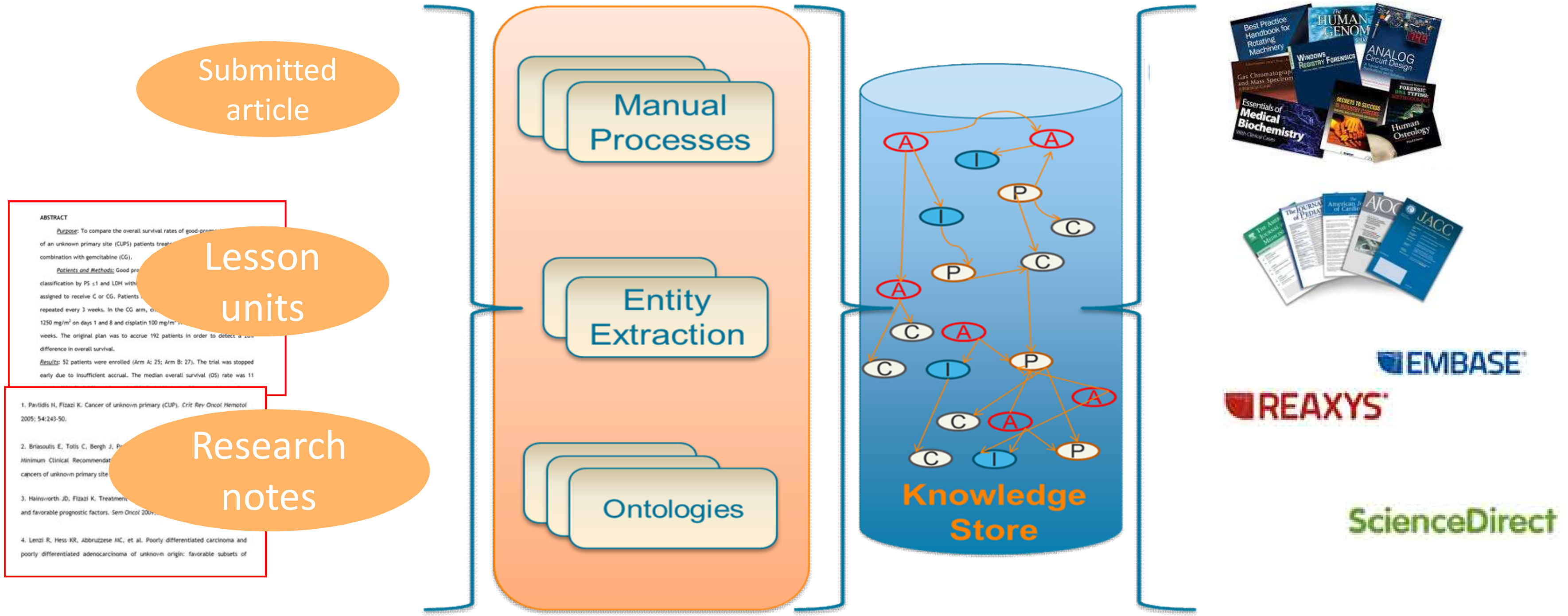
- **Proof-of-concept work at Elsevier Labs built in 2015**
- **Unsupervised, scalable and built with off-the-shelf technologies**
- **Based on recent work at University College London**
  - Riedel, Sebastian, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. "Relation extraction with matrix factorization and universal schemas." (2013).



KNOWLEDGE GRAPH - CREATION

- **Elsevier has the data and core structures to fill a knowledge graph –**
  - Semantic Models / Taxonomies
  - Enrichment pipelines
    - Relation and Fact extraction is currently poor but in progress
    - SD Books initiative: extract definitions
  - Open territory:
    - Current glossaries in books
    - Current acknowledgments in books
  - Research Entities: Authors, Institutions, Publications, Journals, …
  - Curated content
    - The right balance between automated processing and hand annotation
    - *Provenance*- proving trusted source can be differentiator for the Elsevier Knowledge Graph
  - Usage Data
    - Co-usage, downloads, popularity ranks

Knowledge Graph – Strategic Fit



ABSTRACT

Purpose: To compare the overall survival rates of good-prog
of an unknown primary site (CUPS) patients treat
combination with gemcitabine (CG).

Patients and Methods: Good pro
classification by PS ≤1 and LDH withi
assigned to receive C or CG. Patients
repeated every 3 weeks. In the CG arm, c
1250 mg/m^2 on days 1 and 8 and cisplatin 100 mg/m^2
weeks. The original plan was to accrue 192 patients in order to detect a 20
difference in overall survival.

Results: 52 patients were enrolled (Arm A: 25; Arm B: 27). The trial was stopped early due to insufficient accrual. The median overall survival (OS) rate was 11

1. Pavlidis N, Fizazi K. Cancer of unknown primary (CUP). *Crit Rev Oncol Hemotol* 2005; 54:243-50.

2. Briasoulis E, Tolis C, Bergh J, P
Minimum Clinical Recommendat
cancers of unknown primary site

3. Hainsworth JD, Fizazi K. Treatment
and favorable prognostic factors. *Sem Oncol* 200

4. Lenzi R, Hess KR, Abbruzzese MC, et al. Poorly differentiated carcinoma and poorly differentiated adenocarcinoma of unknown origin: favorable subsets of

A single enrichment framework applied to all data types

Common, flexible data structures

Products can make use of multiple enrichment capabilities

Realign content with a Digital Assets model: books and journals *are just one possible rendering of the content*

KNOWLEDGE GRAPH – USES AND APPLICATIONS

- **Flexible disambiguation of entities**
  - Authors, Institutes, Concepts, -- any entity
  - For enrichment pipelines, reference to a knowledge graph with rich data associated with entities will help resolve entities. Enrich entities from:
    - Taxonomies, Wikipedia, DbPedia, Elsevier Sources
  - Powered by existing associations in the graph
- **Query Expansion**
  - Query parsing and interpretation (AskReaxys)
  - Faceting search
    - Suggest associated terms – association of many types (Co-occurrence, taxonomic relations, text-based relations)
- **Recommendations**
  - SD Books use case: background reading
  - Social: often-read together, …
- **Content Generation and Presentation**
  - Question creation
  - Summarization
  - Reasoning: inferred paths (Gene, Physiology, Chemical, Disease)

CHALLENGE: BEYOND PROOF OF CONCEPT — KNOWLEDGE GRAPHS

- **Construction**
  - What are the productive systems building Knowledge Graph from full-text, full feature articles and patents?
  - What modelling and structuring tooling represents the state-of-the-art in Knowledge Graph creation
  - What evidence is there to show something is state-of-the-art?
- **Valorization**
  - What does the knowledge graph offer that we can't create of higher quality in another way?
  - Ultimate measure is the business value. How can we quantify ROI?
  - What productive instances are there as product offerings – currently – in the space of health, science and technology
  - What could you create to differentiate from the current offerings?

MACHINE AIDED TRANSLATION

- **Elsevier manually translates all of the assets that need translation:**
  - Books
  - Medical References
  - Clinical Products
- **Problems:**
  - The costs of translation is inhibitive
  - The turn-around time for full text translations is huge: 1-2 years.
  - Machine aided translation only goes to a certain point
- **Elsevier owns translated taxonomies, e.g. English-French-Spanish medical taxonomy EMMeT**
- **Challenge:**
  - How can we connect taxonomies to machine aided translation,
  - How much effort is required to link taxonomies to a translation memory.
  - To control consistency of target language terminology
- **Are there off-the-shelf/ specific/generic methods**
  - Generalizable
  - What are the best machine translation offerings that integrate and conform with Elsevier's multilingual assets
- **Are there off-the-shelf taxonomy translation products**
  - Proven in the market