Sharing Ontologies Globally To Speed Science And Healthcare Solutions – OntoPortal

International Ontology Sharing Is Becoming A Reality

A consortium of researchers recently formed an organization dedicated to standardizing how scientists define their ontologies, which are essential for retrieving datasets as well as understanding and reproducing research. The group called OntoPortal Alliance is creating a public repository of internationally shared domain-specific ontologies. All the repositories will be managed with a common OntoPortal appliance that has been tested with AllegroGraph Semantic Knowledge Graph software. This enables any OntoPortal adopter to get all the power, features, maintainability, and support benefits that come from using a widely adopted, state-of-the-art semantic knowledge graph database.

Read the full article at HealthIT Outcomes –

As Dr. Jans Aasman, CEO of Franz Inc. explains, "When building a Knowledge Graph as your enterprise's single source of truth, it's critical to include ontologies and taxonomies. AI applications and complex reasoning analytics require information from both databases and knowledge bases that contain domain information, taxonomies, and ontologies to solve complex questions. To make this possible, we developed a novel hybrid sharding technology called FedShard, which facilitates the combination of data and knowledge required by applications like Montefiore's PALM. But this approach is not unique or specific to Healthcare, it is applicable in many other industries, which is why we are excited about OntoPortal’s plans to bring sharing of domain ontologies to a broad audience.”

Knowledge Graphs: A Single Source of Truth for the Enterprise

Dr. Jans Aasman, CEO, Franz Inc.

The notion of a “single source of truth” for the enterprise has been the proverbial moving goalpost for generations of CIOs, only to be waylaid by brittle technology and unending legacy systems. Truth-seeking visions rebuffed by technological trends have continuously confounded business units trying to achieve growth and market penetration. But technology innovation has finally led us to a point where CIOs can now deliver that truth.

Graphing the Truth

Knowledge graphs possess the power to deliver a single source of truth by linking together any assortment of data sources required, standardizing their diversity of data elements, and eliminating silos. They support the most advanced analytics options and decentralized transactions, which is why they’re now deployed as systems of records for some of the most significant, mission-critical use cases affecting our population.

Because they scale to include almost any number of applications — and link to other knowledge graphs as well — these repositories are the ideal solution for real-time information necessary to inform business users’ performances with concrete, data-supported facts. Most importantly, users can get an exhaustive array of touchpoints pertaining to any customer, product, or interaction with an organization from the knowledge graph, making it a single source of truth.

Read the full article at Dataversity.

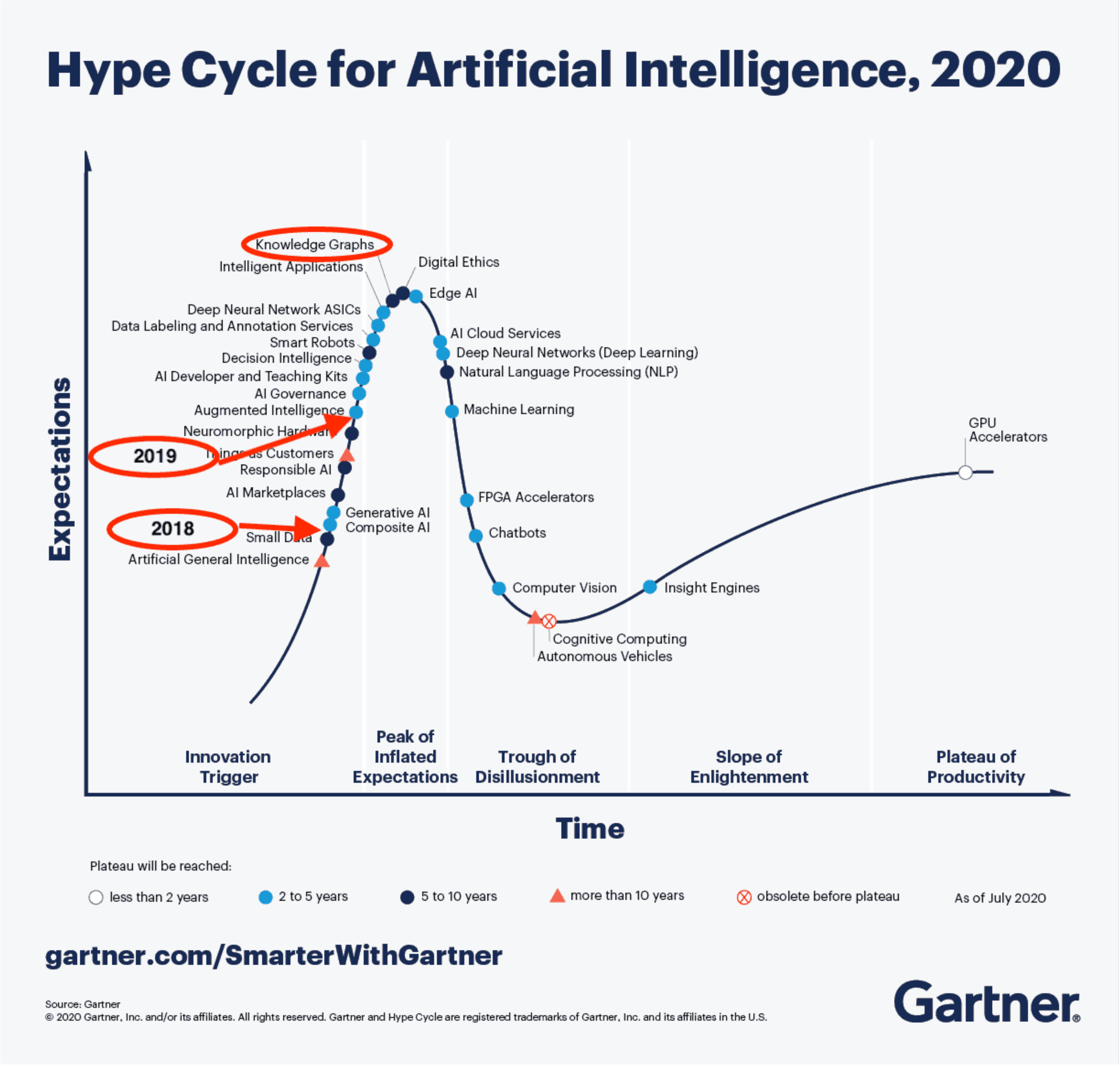
Gartner Hype Cycle for AI – Knowledge Graphs

According to Gartner’s 2020 Hype Cycle for Artificial Intelligence – Despite the global impact of COVID-19, 47% of artificial intelligence (AI) investments were unchanged since the start of the pandemic and 30% of organizations actually planned to increase such investments, according to a Gartner poll. Only 16% had temporarily suspended AI investments, and just 7% had decreased them.

“AI is starting to deliver on its potential and its benefits for businesses are becoming a reality”

Gartner’s – AI Hype Cycle Article

The Hype Cycle growth is consistent with Franz’s customer interest in our Enterprise Knowledge Graph Solutions – Read our recent White Paper.



Connected Data London – The

Future of AI in the Enterprise

The Future of AI in the Enterprise:

Entity-Event Knowledge Graphs for Data-Centric Organizations

Presented by: Dr. Jans Aasman

Register: https://enterprise-kg-cdl-online-meetup.heysummit.com/

Personalized medicine. Predictive call centers. Digital twins for IoT. Predictive supply chain management, and domain-specific Q&A applications. These are just a few AI-driven applications organizations across a broad range of industries are deploying.

Graph databases and Knowledge Graphs are now viewed as a must-have by Enterprises serious about leveraging AI and predictive analytics within their organization.

See how Franz Inc. is helping organizations deploy novel Entity-Event Knowledge Graph Solutions to gain a holistic view of customers, patients, students or other important entities, and the ability to discover deep connections, uncover new patterns and attain explainable results.

Description:

To support ubiquitous AI, a Knowledge Graph system will have to fuse and integrate data, not just in representation, but in context (ontologies, metadata, domain knowledge, terminology systems), and time (temporal relationships between components of data). Building from ‘Entities’ (e.g. Customers, Patients, Bill of Materials) requires a new data model approach that unifies typical enterprise data with knowledge bases such as industry terms and other domain knowledge.

Entity-Event Knowledge Graphs are about connecting the many dots, from different contexts and throughout time, to support and recommend industry-specific solutions that can take into account all the subtle differences and nuisances of entities and their relevant interactions to deliver insights and drive growth. The Entity-Event Data Model we present puts core entities of interest at the center and then collects several layers of knowledge related to the entity as 'Events'.

Franz Inc. is working with organizations across a broad range of industries to deploy large-scale, high-performance Entity-Event Knowledge Graphs that serve as the foundation for AI-driven applications for personalized medicine, predictive call centers, digital twins for IoT, predictive supply chain management and domain-specific Q&A applications—just to name a few.

During this presentation we will explain and demonstrate how Entity-Event Knowledge Graphs are the future of AI in the Enterprise.

Franz Inc. Named an AI 50 Company by KMWorld

AllegroGraph Powering Intelligent Knowledge Graph Solutions

Franz Inc., an early innovator in Artificial Intelligence (AI) and leading supplier of Semantic Graph Database technology for Knowledge Graph Solutions, today announced that it has been named to The AI 50 – Companies Empowering Intelligent Knowledge Management Companies by KMWorld. The annual list reflects the urgency felt among many organizations to provide a timely flow of targeted information. Among the more prominent initiatives is the use of AI and cognitive computing, as well as related capabilities such as machine learning, natural language processing, and text analytics. This list recognizes companies based on their presence, execution, vision and innovation in delivering products and services to the marketplace.

"As the drive for digital transformation becomes an imperative for companies seeking to compete and succeed in all industry sectors, intelligent tools and services are being leveraged to enable speed, insight, and accuracy," said Tom Hogan, Group Publisher at KMWorld. "To showcase organizations that are incorporating AI and an assortment of related technologies—including natural language processing, machine learning, and computer vision—into their offerings, KMWorld created the "AI 50: The Companies Empowering Intelligent Knowledge Management."

"Franz Inc. has a rich history in AI and we are honored to receive this acknowledgement for our efforts in delivering AI Knowledge Graph Solutions," said Dr. Jans Aasman, CEO, Franz Inc. "In the past year, we have seen demand for Intelligent Data Fabrics take off across industries along with recognition from top technology analyst firms that Knowledge Graphs provide the critical foundation for Enterprise Wide Data Fabrics. Our recent launch of AllegroGraph 7 with FedShard, a breakthrough that allows infinite data integration to unify all data and siloed knowledge into an Entity-Event Knowledge Graph solution will catalyze Data Fabric deployments across the Enterprise."

Gartner's Top 10 Trends in Data and Analytics for 2020 noted "Relationships form the foundation of data and analytics value. By 2023, graph technologies will facilitate rapid contextualization for decision making in 30% of organizations worldwide. Graph analytics is a set of analytic techniques that allows for the exploration of relationships between entities of interest such as organizations, people and transactions. Data and analytics leaders need to evaluate opportunities to incorporate graph analytics into their analytics portfolios and applications to uncover hidden patterns and relationships. In addition, consider investigating how graph algorithms and technologies can improve your AI and ML initiatives." (Source: Gartner, Top 10 Trends in Data and Analytics for 2020, June 9, 2020).

"Graph databases and knowledge graphs are now viewed as a must-have by enterprises serious about leveraging AI and predictive analytics within their organization," said Dr. Aasman "We are working with organizations across a broad range of industries to deploy large-scale, high-performance Entity-Event Knowledge Graphs that serve as the foundation for AI-driven Data Fabrics for personalized medicine, predictive call centers, digital twins for IoT, predictive supply chain management and domain-specific Q&A applications – just to name a few."

Forrester Shortlists AllegroGraph

AllegroGraph was shortlisted in the February 3, 2020 Forrester Now Tech: Graph Data Platforms, Q1 2020 report, which recommends that organizations "Use graph data platforms to accelerate connected-data initiatives." Forrester states, "You can use graph data platforms to become significantly more productive, deliver accurate customer recommendations, and quickly make connections to related data."

Bloor Research covers AllegroGraph with FedShard

Bloor Research Analyst, Daniel Howard noted "With the 7.0 release of AllegroGraph, arguably the most compelling new capability is its ability to create what Franz refers to as "Entity-Event Knowledge Graphs" (or EEKGs) via its patented FedShard technology." Mr. Howard goes on to state "Franz clearly considers this a major release for AllegroGraph. Certainly, the introduction of an explicit entity-event graph is not something I've seen before. The newly introduced text to speech capabilities also seem highly promising."

AllegroGraph Named to DBTA's 100 Companies That Matter Most in Data

AllegroGraph was also recently named to DBTA's 100 Companies That Matter Most in Data. The DBTA 100 showcases organizations that delivering solutions for customers to meet the need for real-time, data-driven insights.

Franz Knowledge Graph Technology and Services

Franz's Knowledge Graph Solution includes both technology and services for building industrial strength Entity-Event Knowledge Graphs based on best-of-class tools, products, knowledge, skills and experience. At the core of the solution is Franz's graph database technology, AllegroGraph with FedShard, which is utilized by dozens of the top F500 companies worldwide and enables businesses to extract sophisticated decision insights and predictive analytics from highly complex, distributed data that cannot be uncovered with conventional databases.

Franz delivers the expertise for designing ontology and taxonomy-based solutions by utilizing standards-based development processes and tools. Franz also offers data integration services from siloed data using W3C industry standard semantics, which can then be continually integrated with information that comes from other data sources. In addition, the Franz data science team provides expertise in custom algorithms to maximize data analytics and uncover hidden knowledge.

About Franz Inc.

Franz Inc. is an early innovator in Artificial Intelligence

(AI) and leading supplier of Semantic Graph Database technology with expert knowledge in developing and deploying Knowledge Graph solutions. The foundation for Knowledge Graphs and AI lies in the facets of semantic technology provided by AllegroGraph with FedShard and Allegro CL. The ability to rapidly integrate new knowledge is the crux of the Knowledge Graph and Franz Inc. provides the key technologies and services to address your complex challenges. Franz Inc. is your Knowledge Graph technology partner.

All trademarks and registered trademarks in this document are the properties of their respective owners.

AllegroGraph Named to 100 Companies That Matter Most in Data

Franz Inc. Acknowledged as a Leader for Knowledge Graph Solutions

Lafayette, Calif., June 23, 2020 — Franz Inc., an early innovator in Artificial Intelligence (AI) and leading supplier of Semantic Graph Database technology for Knowledge Graph Solutions, today announced that it has been named to The 100 Companies That Matter in Data by Database Trends and Applications. The annual list reflects the urgency felt among many organizations to provide a timely flow of targeted information. Among the more prominent initiatives is the use of AI and cognitive computing, as well as related capabilities such as machine learning, natural language processing, and text analytics.   This list recognizes companies based on their presence, execution, vision and innovation in delivering products and services to the marketplace.

“We’re excited to announce our eighth annual list, as the industry continues to grow and evolve,” remarked Thomas Hogan, Group Publisher at Database Trends and Applications. “Now, more than ever, businesses are looking for ways transform how they operate and deliver value to customers with greater agility, efficiency and innovation. This list seeks to highlight those companies that have been successful in establishing themselves as unique resources for data professionals and stakeholders.”

“We are honored to receive this acknowledgement for our efforts in delivering Enterprise Knowledge Graph Solutions,” said Dr. Jans Aasman, CEO, Franz Inc. “In the past year, we have seen demand for Enterprise Knowledge Graphs take off across industries along with recognition from top technology analyst firms that Knowledge Graphs provide the critical foundation for artificial intelligence applications and predictive analytics.

Our recent launch of AllegroGraph 7 with FedShard, a breakthrough that allows infinite data integration to unify all data and siloed knowledge into an Entity-Event Knowledge Graph solution will catalyze Knowledge Graph deployments across the Enterprise.”

Gartner recently released a report “How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications” and have previously stated, “The application of graph processing and graph databases will grow at 100 percent annually through 2022 to continuously accelerate data preparation and enable more complex and adaptive data science.” To that end, Gartner named graph analytics as a “Top 10 Data and Analytics Trend” to solve critical business priorities. (*Source: Gartner, Top 10 Data and Analytics Trends, November 5, 2019)*.

“Graph databases and knowledge graphs are now viewed as a must-have by enterprises serious about leveraging AI and predictive analytics within their organization," said Dr. Aasman "We are working with organizations across a broad range of industries to deploy large-scale, high-performance Entity-Event Knowledge Graphs that serve as the foundation for AI-driven applications for personalized medicine, predictive call centers, digital twins for IoT, predictive supply chain management and domain-specific Q&A applications – just to name a few."

Forrester Shortlists AllegroGraph

AllegroGraph was shortlisted in the February 3, 2020 Forrester Now Tech: Graph Data Platforms, Q1 2020 report, which recommends that organizations "Use graph data platforms to accelerate connected-data initiatives." Forrester states, "You can use graph data platforms to become significantly more productive, deliver accurate customer recommendations, and quickly make connections to related data."

Bloor Research covers AllegroGraph with FedShard

Bloor Research Analyst, Daniel Howard noted "With the 7.0 release of AllegroGraph, arguably the most compelling new capability is its ability to create what Franz refers to as "Entity-Event Knowledge Graphs" (or EEKGs) via its patented FedShard technology." Mr. Howard goes on to state "Franz clearly considers this a major release for AllegroGraph. Certainly, the introduction of an explicit entity-event graph is not something I've seen before. The newly introduced text to speech capabilities also seem highly promising."

AllegroGraph Named to KMWorld's 100 Companies That Matter in Knowledge Management

AllegroGraph was also recently named to KMWorld's 100 Companies That Matter in Knowledge Management. The KMWorld 100 showcases organizations that are advancing their products and capabilities to meet changing requirements in Knowledge

Management.

Franz Knowledge Graph Technology and Services

Franz’s Knowledge Graph Solution includes both technology and services for building industrial strength Entity-Event Knowledge Graphs based on best-of-class tools, products, knowledge, skills and experience. At the core of the solution is Franz’s graph database technology, AllegroGraph with FedShard, which is utilized by dozens of the top F500 companies worldwide and enables businesses to extract sophisticated decision insights and predictive analytics from highly complex, distributed data that cannot be uncovered with conventional databases.

Franz delivers the expertise for designing ontology and taxonomy-based solutions by utilizing standards-based development processes and tools. Franz also offers data integration services from siloed data using W3C industry standard semantics, which can then be continually integrated with information that comes from other data sources. In addition, the Franz data science team provides expertise in custom algorithms to maximize data analytics and uncover hidden knowledge.

Ubiquitous AI Demands A New Type Of Database Sharding

Forbes published the following article by Dr. Jans Aasman, Franz Inc.’s CEO.

The notion of sharding has become increasingly crucial for selecting and optimizing database architectures. In many cases, sharding is a means of horizontally distributing data; if properly implemented, it results in near-infinite scalability. This option enables database availability for business continuity, allowing organizations to replicate databases among geographic locations. It’s equally useful for load balancing, in which computational necessities (like processing) shift between machines to improve IT resource allocation.

However, these use cases fail to actualize sharding’s full potential to maximize database performance in today’s post-big data landscape. There’s an even more powerful form of sharding, called “hybrid sharding,” that drastically improves the speed of query results and duly expands the complexity of the questions that can be asked and answered. Hybrid sharding is the ability to combine data that can be partitioned into shards with data that represents knowledge that is usually un-shardable.

This hybrid sharding works particularly well with the knowledge graph phenomenon leveraged by the world’s top data-driven companies. Hybrid sharding also creates the enterprise scalability to query scores of internal and external sources for nuanced, detailed results, with responsiveness commensurate to that of the contemporary AI age.

Read the full article at Forbes.

NEW! – Franz’s AllegroGraph 7 Powers First Distributed Semantic Knowledge Graph Solution with Federated-Sharding

FedShard™, Entity-Event Data Modeling and Browser-based Gruff Drives Infinite Data Integration, Holistic Insights and Complex Reasoning

Franz Inc., an early innovator in Artificial Intelligence (AI) and leading supplier of Semantic Graph Database technology for Knowledge Graph Solutions, today announced AllegroGraph 7, a breakthrough solution that allows infinite data integration through a patented approach unifying all data and siloed knowledge into an Entity-Event Knowledge Graph solution that can support massive big data analytics. AllegroGraph 7 utilizes unique federated sharding capabilities that drive 360-degree insights and enable complex reasoning across a distributed Knowledge Graph. Hidden connections in data are revealed to AllegroGraph 7 users through a new browser-based version of Gruff, an advanced visualization and graphical query builder.

“Large enterprises have Knowledge Graphs that are so big that no amount of vertical scaling will work,” said Jans Aasman, CEO of Franz Inc. “When these organizations want to conduct new big data analytics, it requires a new effort by the IT department to gather semi-usable data for the data scientists, which can cost millions of dollars, waste valuable time and still not provide a holistic data architecture for querying across all data. ETL, Data Lakes and Property Graphs only exacerbate the problem by creating new data silos. AllegroGraph 7 takes a holistic approach to mixed data, unifying all enterprise data with domain knowledge, including taxonomies, ontologies and industry knowledge – making queries across all data possible, while simplifying and accelerating feature extraction for machine learning."

To support ubiquitous AI, a Knowledge Graph system will have to fuse and integrate data, not just in representation, but in context (ontologies, metadata, domain knowledge, terminology systems), and time (temporal relationships between components of data). The rich functional and contextual integration of multi-modal, predictive modeling and artificial intelligence is what distinguishes AllegroGraph 7 as a modern, scalable, enterprise analytic platform. AllegroGraph 7 is the first big temporal knowledge graph technology that encapsulates a novel entity-event model natively integrated with domain ontologies and metadata, and dynamic ways of setting the analytics lens on all entities in the system (patient, person, devices, transactions, events, and operations) as prime objects that can be the focus of an analytic (AI, ML, DL) process.

AI applications and complex reasoning analytics require information from both databases and knowledge bases that contain domain information, taxonomies and ontologies in order to conduct queries. Some large-scale knowledge bases cannot be sharded because they contain highly interconnected data. AllegroGraph 7 federates any shard with any large-scale knowledge base – providing a novel way to shard knowledge bases without duplicating knowledge bases in every shard. This approach creates a modern analytic system that integrates data in context (ontologies, metadata, domain knowledge, terminology systems) and time (temporal relationships between components of data). The result is a rich functional and contextual integration of data suitable for large scale analytics, predictive modeling, and artificial intelligence.

Financial institutions, healthcare providers, contact centers, manufacturing firms, government agencies and other large enterprises that use AllegroGraph 7 gain a holistic, future-proofed Knowledge Graph architecture for big data predictive analytics and machine learning across complex knowledge bases.

“AllegroGraph 7’s support of Entity-Event Data Modeling is the most welcome innovation and addition to our arsenal in reimagining healthcare and implementing Precision Medicine,” said Dr. Parsa Mirhaji, Director of Center for Health Data Innovations at the Albert Einstein College of Medicine and Montefiore Health System, NY “Precision Medicine is about moving away from statistical averages and broad-based patterns. It is about connecting many dots, from different contexts and throughout time, to support precision diagnosis and to recommend the precision care that can take into account all the subtle differences and nuisances of individuals and their personal experiences throughout their life. This technology is about saving lives, by leveraging data, context and analytics and is what Franz’s Entity-Event Data Modeling brings to the table.”

Dr. Mirhaji and his team at Montefiore Health System have developed the Patient-centered Analytic Learning Machine (PALM) using these capabilities to provide an enterprise platform for Artificial Intelligence and machine learning in healthcare that can support conversational AI, interpret data from EMR, natural language, and radiological images, all centered around life-time experiences of an individual patient. A single system that unifies all analytics and data from heterogeneous sources to manage appointments and prescriptions, triage patients with potential spinal cancer, respiratory failure, or sepsis, and provide just-in-time recommendations and personalized decision support for clinicians to improve patients’ outcomes.

Key capabilities in AllegroGraph 7 include:

Semantic Entity-Event Data Modeling

Big Data predictive analytics requires a new data model approach that unifies typical enterprise data with knowledge bases such as taxonomies, ontologies, industry terms and other domain knowledge. The Entity-Event Data Model utilized by AllegroGraph 7 puts core ‘entities’ such as customers, patients, students or people of interest at the center and then collects several layers of knowledge related to the entity as ‘events.’ The events represent activities that transpire in a temporal context. Using this novel data model approach, organizations gain a holistic view of customers, patients, students or important entities and the ability to discover deep connections, uncover new patterns and attain explainable results.

FedShard™ Speeds Complex Queries

Through a patented in-memory federation function, the results from each machine are combined so that the query process appears as if only one database is being accessed, although many different databases and data stores and knowledge bases are actually being accessed and returning results. This unique data federation capability accelerates results for highly complex queries across highly distributed data sets and knowledge bases.

Large-scale Mixed Data Processing

The AllegroGraph 7 big data processing system is able to scale massive amounts of domain knowledge data by efficiently associating domain knowledge with partitioned data through shardable graphs on clusters of machines. AllegroGraph 7 efficiently combines partitioned data with domain knowledge through an innovative process that keeps as much of the data in RAM as possible to speed data access and fully utilize the processors of the query servers.

Browser-based Gruff
Gruff's powerful query and visualization capabilities are now available via a web browser and directly integrated in AllegroGraph 7. Gruff is the industry's leading Knowledge Graph visualization tool that dynamically displays visual graphs and related links. Gruff's 'Time Machine' provides users with an important capability to explore temporal connections and see how relationships are created over time. Users can build visual graphs that display the relationships in graph databases, display tables of properties, manage queries, connect to SPARQL Endpoints, and build SPARQL or Prolog queries as visual diagrams. Gruff can be downloaded separately or is included with the AllegroGraph v7 distribution.

High Performance Big Data Analytics

AllegroGraph 7 delivers high performance analytics by overcoming data processing issues related to disk versus memory access, uses processor core efficiency and updates domain knowledge databases across partitioned data systems in a highly efficient manner.

Gartner predicts "the application of graph processing and graph DBMSs will grow at 100 percent annually through 2022 to continuously accelerate data preparation and enable more complex and adaptive data science." In addition, Gartner named graph analytics as a "Top 10 Data and Analytics Trend" to solve critical business priorities." (*Source: Gartner, Top 10 Data and Analytics Trends, November 5, 2019)*

AllegroGraph 7 Availability

AllegroGraph 7 is immediately available directly from Franz Inc. Visit the AllegroGraph YouTube channel to see AllegroGraph in action.

Join AllegroGraph 7 Webinar
Franz Inc. will host a webcast entitled "Scalable Knowledge

Graphs Using the New Distributed AllegroGraph 7." Register for the Webinar.

Knowledge Graph Conference – May 4 – 7, 2020

Dr. Jans Aasman, CEO, Franz Inc., will be presenting a talk at the Knowledge Graph Conference entitled, "The Knowledge Graph that Listens" on May 7th at 1PM Eastern. Register for the Conference.

The Knowledge Graph Cookbook

Released April 22, 2020, this new book directs readers on why and how to build Knowledge Graphs that help enterprises use data to innovate, create value and increase revenue. The book is full of recipes and knowledge on the subject and features an interview with Dr. Jans Aasman, CEO, Franz Inc. in the Expert Opinion section. Get a copy of the book.

Natural Language Processing and Machine Learning in AllegroGraph

The majority of our customers build Knowledge Graphs with Natural Language and Machine learning components. Because of this trend AllegroGraph now offers strong support for the use of Natural Language Processing and Machine learning.

Franz Inc has a team of NLP engineers and Taxonomy experts that can help with building turn-key solutions. In general however, our customers already have some expertise in house. In those cases we train customers in how to take the output of NLP and ML processing and turn that into an efficient Knowledge Graph based on best practices in the industry.

This document primarily describes the NLP and ML plug-in AllegroGraph.

Note that many enterprises already have a data science team with NLP experts that use modern open source NLP tools like Spacy, Gensim or Polyglot, or Machine Learning based NLP tools like BERT and Scikit-Learn. In another blog about Document Handling we describe a pipeline of how to deal with NLP in Document Knowledge Graphs by using our NLP and ML plugin and mix that with open source tools.

PlugIn features for Natural Language Processing and Machine Learning in AllegroGraph.

Here is the outline of the plugin features that we are going to describe in more detail.

Machine learning

- data acquisition
- classifier training
- feature extraction support
- performance analysis
- model persistence

NLP

- handling languages
- handling dictionaries
- tokenization
- entity extraction
- Sentiment analysis
- basic pattern matching

SPARQL Access

- Future development

Machine Learning

ML: Data Acquisition

Given that the NLP and ML functions operate within AllegroGraph, after loading the plugins, data acquisition can be performed directly from the triple-store, which drastically simplifies the data scientist workflow. However, if the data is not in AllegroGraph yet we can also import it directly from ten formats of triples or we can use our additional capabilities to import from CSV/JSON/JSON-LD.

Part of the Data Acquisition is also that we need to pre-process the data for training so we provide these three functions:

- prepare-training-data
- split-dev-test
- equalize (for resampling)

Machine Learning: Classifiers

- Currently we provide simple linear classifiers. In case there’s a need for neural net or other advanced classifiers, those can be integrated on-demand.
- We also provide support for online learning (online machine learning is an ML method in which data becomes available in a sequential order and is used to update the best predictor for future data at each step, as opposed to batch learning techniques which generate the best predictor by learning on the entire training data set at once). This feature is useful for many real-world data sets that are constantly updated.
- The default classifiers available are Averaged

Machine Learning: Feature Extraction

Each classifier is expecting a vector of features: either feature indices (indicative features) or pairs of numbers (index – value). These are obtained in a two-step process:

1. A classifier-specific extract-features method should be defined that will return raw feature vector with features identified by strings of the following form: prefix|feature.

The prefix should be provided as a keyword argument to the collect-features method call, and it is used to distinguish similar features from different sources (for instance, for distinct predicates).

2. Those features will be automatically transformed to unique integer ids. The resulting feature vector of indicator features may look like the following: #(1 123 2999 …)

Note that these features may be persisted to AllegroGraph for repeated re-use (e.g. for experimenting with classifier hyperparameter tuning or different classification models).

Many possible features may be extracted from data, but there is a set of common ones, such as:

1. individual tokens of the text field
2. ngrams (of a specified order) of the text field
3. presence of a token in a specific dictionary (like, the dictionary of slang words)
4. presence/value of a certain predicate for the subject of the current triple
5. length of the text

And in case the user has a need for special types of tokens we can write specific token methods, here is an example (in Lisp) that produces an indicator feature of a presence of emojis in the text:

```
(defmethod collect-features ((method (eql :emoji)) toks &key
pred)
(dolist (tok toks)
(when (some #'(lambda (code)
  (or (<= #x1F600 code #x1F64F)
      (<= #x1F650 code #x1F67F)
      (<= #x1F680 code #x1F6FF)))
   (map 'vector #'char-code tok))
(return (list "emoji")))))
```

Machine Learning: Integration with Spacy

The NLP and ML community invents new features and capabilities at an incredible speed. Way faster than any database company can keep up with. So why not embrace that? Whenever we need something that we don’t have in AllegroGraph yet we can call out to Spacy or any other external NLP tool. Here is an example of using feature extraction from Spacy to collect indicator features of the text dependency parse relations:

```
(defmethod collect-features ((method (eql :dep)) deps &key
pred dep-type dep-labels)
 (loop :for ds :in deps :nconc
  (loop :for dep :in ds
   :when (and (member (dep-tag dep) dep-labels)
              (dep-head dep)
              (dep-tok dep))
    :collect (format nil "dep|~a|~a_~a"
              dep-type
              (tok-word (dep-head dep)
              (tok-word (dep-tok dep))))))
```

The demonstrated integration uses Spacy Docker instance and its HTTP API.

Machine Learning: Classifier Analysis

We provide all the basic tools and metrics for classifier quality analysis:

- accuracy
- f1, precision, recall
- confusion matrix
- and an aggregated classification report

Machine Learning: Model Persistence

The idea behind model persistence is that all the data can be stored in AllegroGraph, including features and classifier models. AllegroGraph stores classifiers directly as triples. This is a far more robust and language-independent approach than currently popular among data scientists reliance on Python pickle files. For the storage we provide a basic triple-based format, so it is also possible to interchange the models using standard RDF data formats.

The biggest advantage of this approach is that when adding text to AllegroGraph we don’t have to move the data externally to perform the classification but can keep the whole pipeline entirely internal.

Natural Language Procession (NLP)

NLP: Language Packs

Most of the NLP tools are language-dependent: i.e. there’s a general function that uses language-specific model/rules/etc. In AllegroGraph, support for particular languages is provided on-demand and all the language-specific is grouped in the so called “language pack” or langpack, for short – a directory with a number of text and binary files with predefined names.

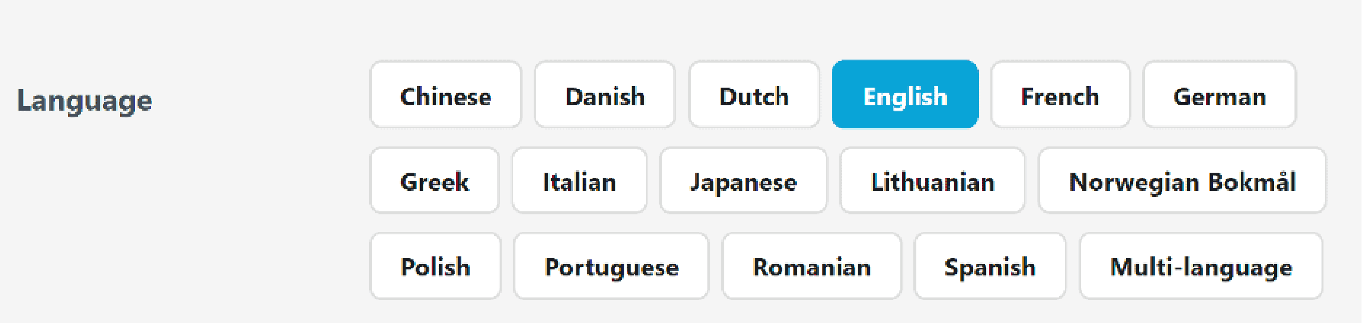
Currently, the langpack for English is provided at nlp/langs/en.zip, with the following files:

- contractions.txt – a dictionary of contractions
- abbrs.txt – a dictionary of abbreviations
- stopwords.txt – a dictionary of stopwords
- pos-dict.txt – positive sentiment words
- neg-dict.txt – negative sentiment words
- word-tok.txt – a list of word tokenization rules

Additionally, we use a general dictionary, a word-form dictionary (obtained from Wiktionary), and custom lexicons.

Loading a langpack for a particular language is performed using load-langpack.

Creating a langpack is just a matter of adding the properly named files to the directory and can be done manually. The names of the files should correspond to the names of the dictionary variables that will be filled by the pack. The dictionaries that don't have a corresponding file will be just skipped.We have just finished creating a langpack for Spanish and it will be published soon. In case you need other dictionaries we use our AG/Spacy infrastructure. Spacy recently added a comprehensive list of new languages:



NLP: Dictionaries

Dictionaries are read from the language packs or other sources and are kept in memory as language-specific hash-tables. Alongside support for storing the dictionaries as text files, there are also utilities for working with them as triples and putting them into the triple store.

Note that we at Franz Inc specialize in Taxonomy Building using various commercial taxonomy building tools. All these tools can now export these taxonomies as a mix of SKOS taxonomies and OWL. We have several functions to read directly from these SKOS taxonomies and turn them into dictionaries that support efficient phrase-level lookup.

NLP: Tokenization

Tokenization is performed using a time-proven rule-based approach. There are 3 levels of tokenization that have both a corresponding specific utility function and an :output format of the tokenize function:

> :parags – splits the text into a list of lists of tokens for paragraphs and sentences in each paragraph
> :sents – splits the text into a list of tokens for each sentence
> :words – splits the text into a plain list of tokens

Paragraph-level tokenization considers newlines as paragraph delimiters. Sentence-level tokenization is geared towards western-style writing that uses dot and other punctuation marks to delimit sentences. It is, currently, hard-coded, but if the need arises, additional handling may be added for other writing systems. Word-level tokenization is performed using a language-specific set of rules.

NLP: Entity Extraction

Entity extraction is performed by efficient matching (exactly or fuzzy) of the token sequences to the existing dictionary structure.

It is expected that the entities come from the triple store and there’s a special utility function that builds lookup dictionaries from all the triples of the repository identified by certain graphs that have a skos:prefLabel or skos:altLabel property. The lookup may be case-insensitive with the exception of abbreviations (default) or case-sensitive.

Similar to entity extraction, there’s also support for spotting sentiment words. It is performed using the positive/negative words dictionaries from the langpack.

One feature that we needed to develop for our customers is ‘heuristic entity extraction’ . In case you want to extract complicated product names from text or call-center conversations between customers and agents you run into the problem that it becomes very expensive to develop altLabels in a taxonomy tool. We created special software to facilitate the automatic creation of altlabels.

NLP: Basic Pattern Matching for relationship and event detection

Getting entities out of text is now well understood and supported by the software community. However, to find complex concepts or relationships between entities or even events is way harder and requires a flexible rule-based pattern matcher. Given our long time background in Lisp and Prolog one can imagine we created a very powerful pattern matcher.

SPARQL Access

Currently all the features above can be controlled as stored procedures or using Lisp as the command language. We have a new (beta) version that uses SPARQL for most of the control. Here are some examples. Note that fai is a magic-property namespace for “AI”-related stuff and inc is a custom namespace of an imaginary client:

1. Entity extraction

```
select ?ent {
   ?subj fai:entityTaxonomy inc:products .
   ?subj fai:entityTaxonomy inc:salesTerms .
   ?subj fai:textPredicate inc:text .
     ?subj fai:entity(fai:language "en", fai:taxonomy
inc:products) ?ent .
}
```

The expressions ?subj fai:entityTaxonomy inc:poducts and ?subj fai:entityTaxonomy inc:salesTerms specify which taxonomies to use (the appropriate matchers are cached).
The expression ?subj fai:entity ?ent will either return the already extracted entities with the specified predicate (fai:entity) or extract the new entities according to the taxonomies in the texts accessible by fai:textPredicate.

2. fai:sentiment will return a single triple with sentiment score:

```
select ?sentiment {
   ?subj fai:textPredicate inc:text .
   ?subj fai:sentiment ?sentiment .
   ?subj fai:language "en" .
   ?subj fai:sentimentTaxonomy franz:sentiwords .
}
```

3. Text classification:
Provided inc:customClassifier was already trained previously, this query will return labels for all texts as a result of classification.

```
select ?label {
?subj fai:textPredicate inc:text .
?subj fai:classifier inc:customClassifier .
?subj fai:classify ?label .
?label fai:storeResultPredicate inc:label .
}
```

Further Development

Our team is currently working on these new features:

- A more accessible UI (python client & web) to facilitate NLP and ML pipelines
- Addition of various classifier models
- Sequence classification support (already implemented for a customer project)
- Pre-trained models shipped with AllegroGraph (e.g. English NER)
- Graph ML algorithms (deepwalk, Google Expander)
- Clustering algorithms (k-means, OPTICS)

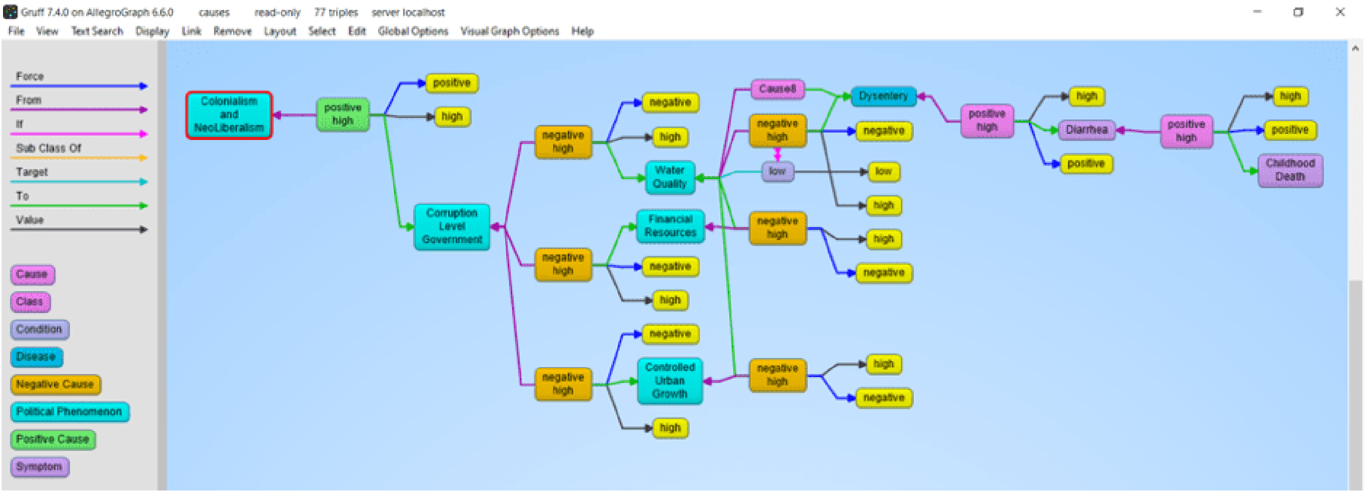
Document Knowledge Graphs with NLP and ML

A core competency for Franz Inc is turning text and documents into Knowledge Graphs (KG) using Natural Language Processing (NLP) and Machine Learning (ML) techniques in combination with AllegroGraph. In this document we discuss how the techniques described in [NLP and ML components of AllegroGraph] can be combined with popular software tools to create a robust Document Knowledge Graph pipeline.

We have applied these techniques for several Knowledge Graphs but in this document we will primarily focus on three completely different examples that we summarize below. First is the Chomsky Legacy Project where we have a large set of very dense documents and very different knowledge sources, Second is a knowledge graph for an intelligent call center where we have to deal with high volume dynamic data and real-time decision support and finally, a large government organization where it is very important that people can do a semantic search against documents and policies that steadily change over time and where it is important that you can see the history of documents and policies.

Example [1] Chomsky Knowledge Graph
The Chomsky Legacy Project is a project run by a group of admirers of Noam Chomsky with the primary goal to preserve all his written work, including all his books, papers and interviews but also everything written about him. Ultimately students, researchers, journalists, lobbyists, people from the AI community, and linguists can all use this knowledge graph for their particular goals and questions.

The biggest challenges for this project are finding causal relationships in his work using event and relationship extraction. A simple example we extracted from an author quoting Chomsky is that neoliberalism ultimately causes childhood death.



Example 2: N3 Results and the Intelligent Call Center
This is a completely different use case (See a recent KMWorld Articlehttps://allegrograph.com/knowledge-graphs-enhance-customer-experience-through-speed-and-accuracy/). Whereas the

previous use case was very static, this one is highly dynamic. We analyze in real-time the text chats and spoken conversations between call center agents and customers. Our knowledge graph software provides real-time decision support to make the call center agents more efficient. N3 Results helps big tech companies to sell their high tech solutions, mostly cloud-based products and services but also helps their clients sell many other technologies and services.

The main challenge we tackle is to really deeply understand what the customer and agent are talking about. None of this can be solved by only simple entity extraction but requires elaborate rule-based and machine learning techniques. Just to give a few examples. We want to know if the agent talked about their most important talking points: that is, did the agent ask if the customer has a budget, or the authority to make a decision or a timeline about when they need the new technology or whether they actually have expressed their need. But also whether the agent reached the right person, and whether the agent talked about the follow-up. In addition, if the customer talks about competing technology we need to recognize that and provide the agent in real-time with a battle card specific to the competing technology. And in order to be able to do the latter, we also analyzed the complicated marketing materials of the clients of N3.

Example 3: Complex Government Documents
Imagine a regulatory body with tens of thousands of documents. Where nearly every paragraph has reference to other paragraphs in the same document or other documents and the documents change over time. The goal here is to provide the end-users in the government with the right document given their current task at hand. The second goal is to keep track of all the changes in the documents (and the relationship between documents) over time.

The Document to Knowledge Graph Pipeline

Process Name	Input	Output
1. Custom Taxonomy Creation	Corpus Analytics, Taxonomy tool	A SKOS taxonomy containing concepts, concept hierarchy, prefLabels, altLabels.
2. Document Preparation	Documents (pdf, word, ppt, xlsx), Apache Tika, Spacy for XML cleanup	An XML version of each document
3. Extract Document Meta Data	Document + Apache Tika	JSON dictionary of the Document MetaData
4. XML-to-Triples	XML+JSON dictionary, XMLToTriples.py	Graph-based document tree with chapters, sections, and paragraphs as triples. Also includes meta data as triples
5. Entity-Extraction	Paragraphs + taxonomies + AllegroGraph Entity extract or external extractors	Concepts, persons, places, currencies. Connected to paragraphs
6. LOD Enrichment	Paragraphs + IBM Natural Language Understanding.	Concept categories and links to DBpedia and GeoNames, etc.
7. Complex Relationship and Event extraction.	Paragraphs + Taxonomy + Rules in Spacy or AllegroGraph	Complex events and relationships, References to other document sections.
8. NLP and ML	Chapters and paragraphs + all the tools described [here], but also using Spacy, Gensim, BERT, SciKit Learn.	Similarities, sentiment, query answering, smart search, text classification, word embeddings, abstracts
9. Versioning and Document tracking	Old + New document, compare.py	Old document in historic repository, new document in current, changed graph.
10. Statistical Relationships	Concepts + OddRatio.py or OddsRatio.cl	Statistical relationships between concepts.

Let us first give a quick summary in words of how we turn documents into a Knowledge Graph.

[1] Taxonomy Creation

Taxonomy of all the concepts important to the business using open source or commercial taxonomy builders. An available industry taxonomy is a good starting point for additional customizations.

[2] Document Preparation

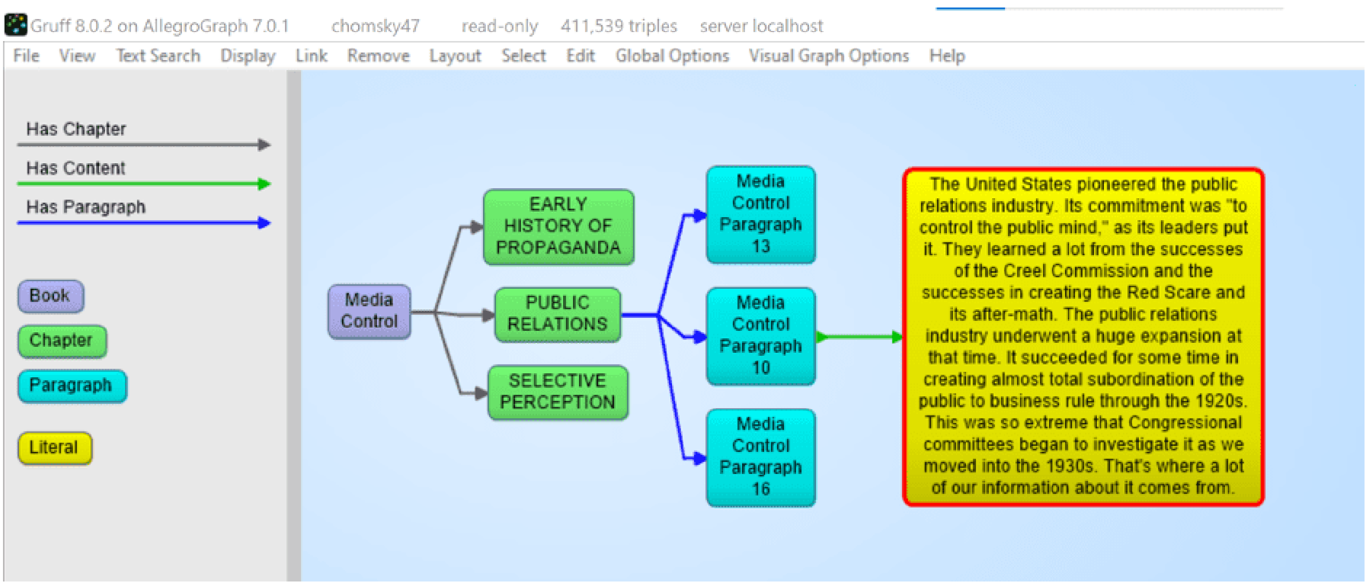
We then take a document and turn it into an intermediate XML using Apache Tika. Apache Tika supports more than 1000 document types and although Apache Tika is a fantastic tool, the output is still usually not clean enough to create a graph from, so we use Spacy rules to clean up the XML to make it as uniform as possible.

[3] Extract Document MetaData

Most documents also contain document metadata (author, date, version, title, etc) and Apache Tika will also deliver the metadata for a document as a JSON object.

[4] XML to Triples

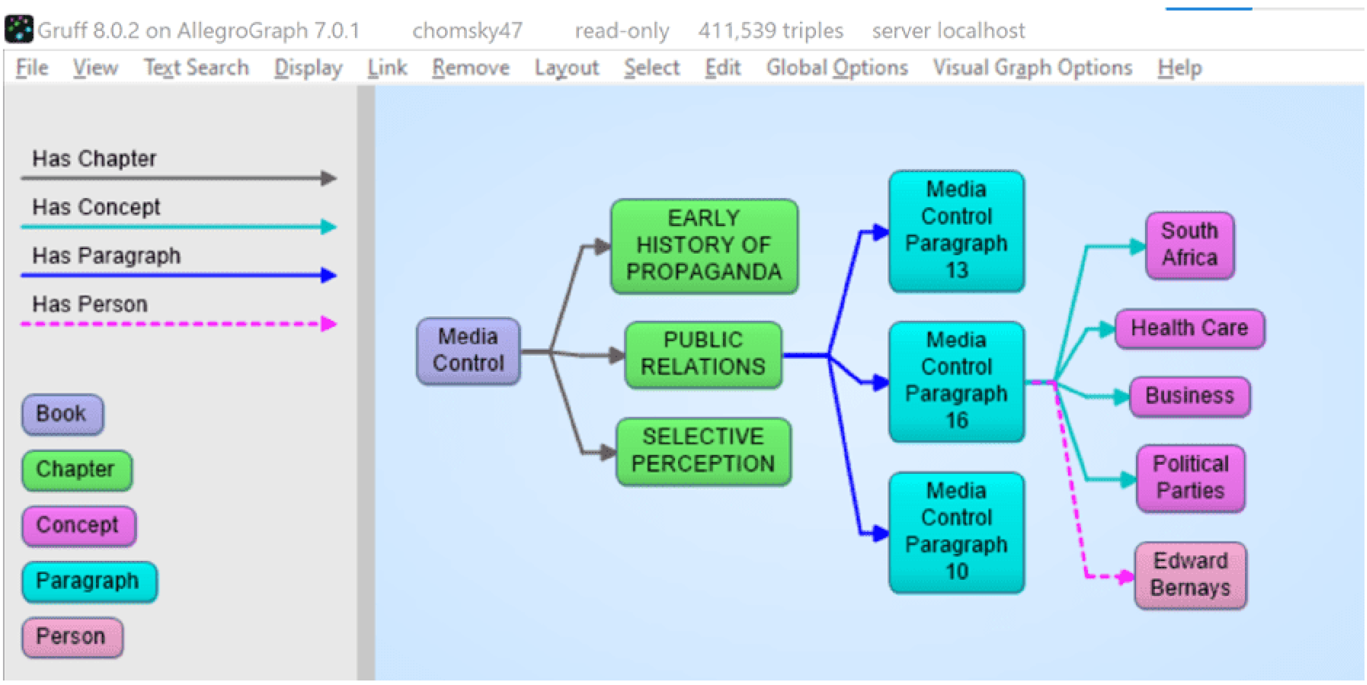
Our tools ingest the XML and metadata and transform that into a graph-based document tree. The document is the root and from that, it branches out into chapters, optionally sections, all the way down to paragraphs. The ultimate text content is in the paragraphs. In the following example we took the XML version of Noam Chomsky’s book Media Control and turned that into a tree. The following shows a tiny part of that tree. We start with the Media Control node, then we show three (of the 11) chapters, for one chapter we show three (of the 6) paragraphs, and then we show the actual text in that paragraph. We sometimes can go even deeper to the level of sentences and tokens but for most projects that is overkill.

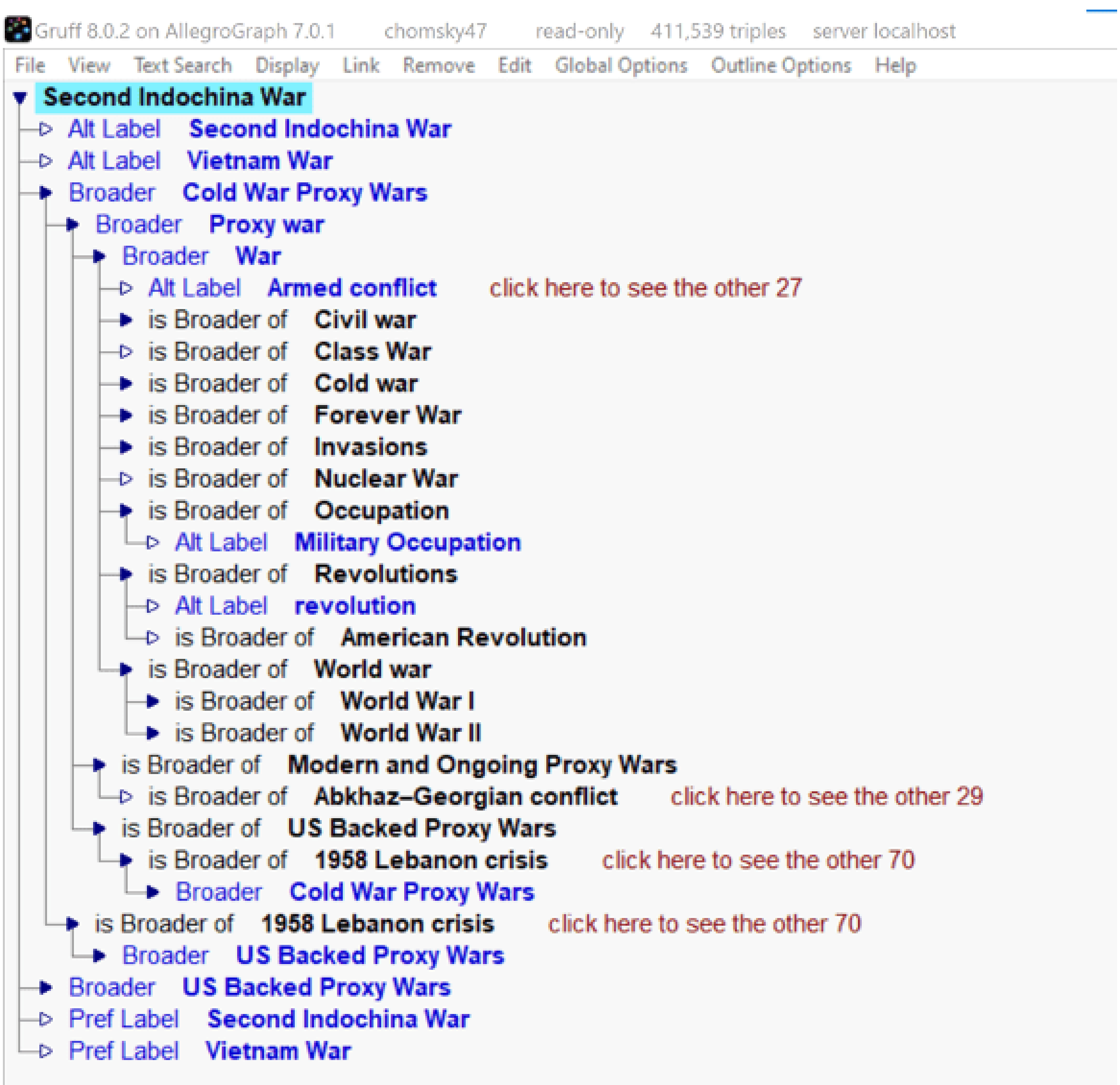


The United States pioneered the public relations industry. Its commitment was "to control the public mind," as its leaders put it. They learned a lot from the successes of the Creel Commission and the successes in creating the Red Scare and its after-math. The public relations industry underwent a huge expansion at that time. It succeeded for some time in creating almost total subordination of the public to business rule through the 1920s. This was so extreme that Congressional committees began to investigate it as we moved into the 1930s. That's where a lot of our information about it comes from.

[5] Entity Extractor

AllegroGraph’s entity extractor takes as input the text of each paragraph in the document tree and one or more of the taxonomies and returns recognized SKOS concepts based on prefLabels and altLabels. AllegroGraph’s entity extractor is state of the art and especially powerful when it comes to complex terms like product names. We find that in our call center a technical product name can sometimes have up to six synonyms or very specific jargon. For example the Cisco product Catalyst 9000 will also be abbreviated as the cat 9k. Instead of developing altLabels for every possible permutation that human beings *will* use, we have specialized heuristics to optimize the yield from the entity extractor. The following picture shows 4 (of the 14) concepts discovered in paragraph 16. Plus one person that was extracted by IBM’s NLU.

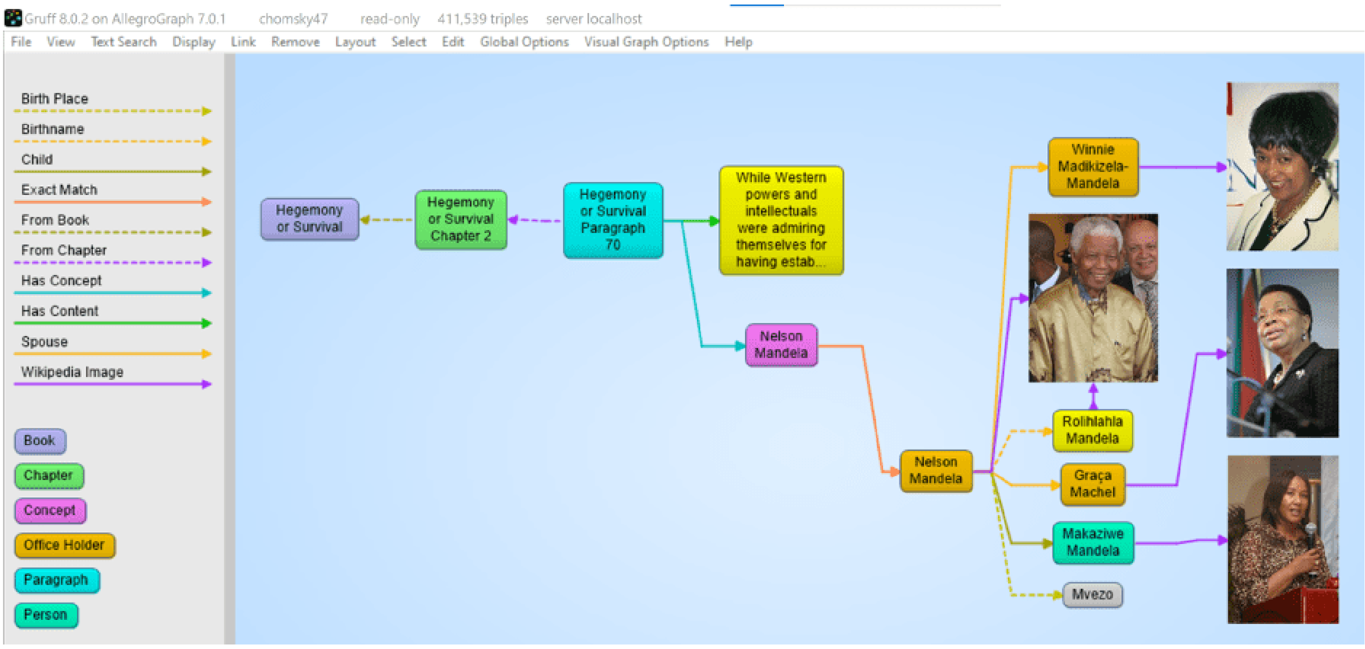

Gruff 8.0.2 on AllegroGraph 7.0.1 chomsky47 read-only 411,539 triples server localhost
File View Text Search Display Link Remove Layout Select Edit Global Options Visual Graph Options Help
Has Chapter
Has Concept
Has Paragraph
Has Person
Book
Chapter
Concept
Paragraph
Person
Media Control
EARLY HISTORY OF PROPAGANDA
PUBLIC RELATIONS
SELECTIVE PERCEPTION
Media Control Paragraph 13
Media Control Paragraph 16
Media Control Paragraph 10
South Africa
Health Care
Business
Political Parties
Edward Bernays


Gruff 8.0.2 on AllegroGraph 7.0.1 chomsky47 read-only 411,539 triples server localhost
File View Text Search Display Link Remove Edit Global Options Outline Options Help

```
▼ Second Indochina War
  ▷ Alt Label   Second Indochina War
  ▷ Alt Label   Vietnam War
  ▶ Broader   Cold War Proxy Wars
    ▶ Broader   Proxy war
      ▶ Broader   War
        ▷ Alt Label   Armed conflict      click here to see the other 27
        ▶ is Broader of   Civil war
        ▷ is Broader of   Class War
        ▶ is Broader of   Cold war
        ▶ is Broader of   Forever War
        ▶ is Broader of   Invasions
        ▷ is Broader of   Nuclear War
        ▶ is Broader of   Occupation
          ▷ Alt Label   Military Occupation
        ▶ is Broader of   Revolutions
          ▷ Alt Label   revolution
          ▷ is Broader of   American Revolution
        ▶ is Broader of   World war
          ▶ is Broader of   World War I
          ▶ is Broader of   World War II
      ▶ is Broader of   Modern and Ongoing Proxy Wars
        ▷ is Broader of   Abkhaz–Georgian conflict      click here to see the other 29
      ▶ is Broader of   US Backed Proxy Wars
        ▶ is Broader of   1958 Lebanon crisis      click here to see the other 70
          ▶ Broader   Cold War Proxy Wars
    ▶ is Broader of   1958 Lebanon crisis      click here to see the other 70
      ▶ Broader   US Backed Proxy Wars
  ▶ Broader   US Backed Proxy Wars
  ▷ Pref Label   Second Indochina War
  ▷ Pref Label   Vietnam War
```

[6] Linked Data Enrichment

In many use cases, AllegroGraph can link extracted entities to concepts in the linked data cloud. The most prominent being DBpedia, wikidata, the census database, GeoNames, but also many Linked Open Data repositories. One tool that is very useful for this is IBM’s Natural Language Understanding program but there are others available. In the following image we see that the Nelson Mandela entity (Red) is linked to the dbpedia entity for Nelson Mandela and that then links to the DBpedia itself. We extracted some of his spouses and a child with their pictures.


Gruff 8.0.2 on AllegroGraph 7.0.1
chomsky47
read-only
411,539 triples
server localhost
File View Text Search Display Link Remove Layout Select Edit Global Options Visual Graph Options Help
Birth Place
Birthname
Child
Exact Match
From Book
From Chapter
Has Concept
Has Content
Spouse
Wikipedia Image
Book
Chapter
Concept
Office Holder
Paragraph
Person
Hegemony or Survival
Hegemony or Survival Chapter 2
Hegemony or Survival Paragraph 70
While Western powers and intellectuals were admiring themselves for having estab...
Nelson Mandela
Nelson Mandela
Winnie Madikizela-Mandela
Rolihlahla Mandela
Graça Machel
Makaziwe Mandela
Mvezo

[7] Complex Relationship and Event Extraction

Entity extraction is a first good step to ‘see’ what is in your documents but it is just the first step. For example: how do you find in a text whether company C1 merged with company C2. There are many different ways to express the fact that a company fired a CEO. For example: Uber got rid of Kalanick, Uber and Kalanick parted ways, the board of Uber kicked out the CEO, etc. We need to write explicit symbolic rules for this or we need a lot of training data to feed a machine learning algorithm.

[8] NLP and Machine Learning

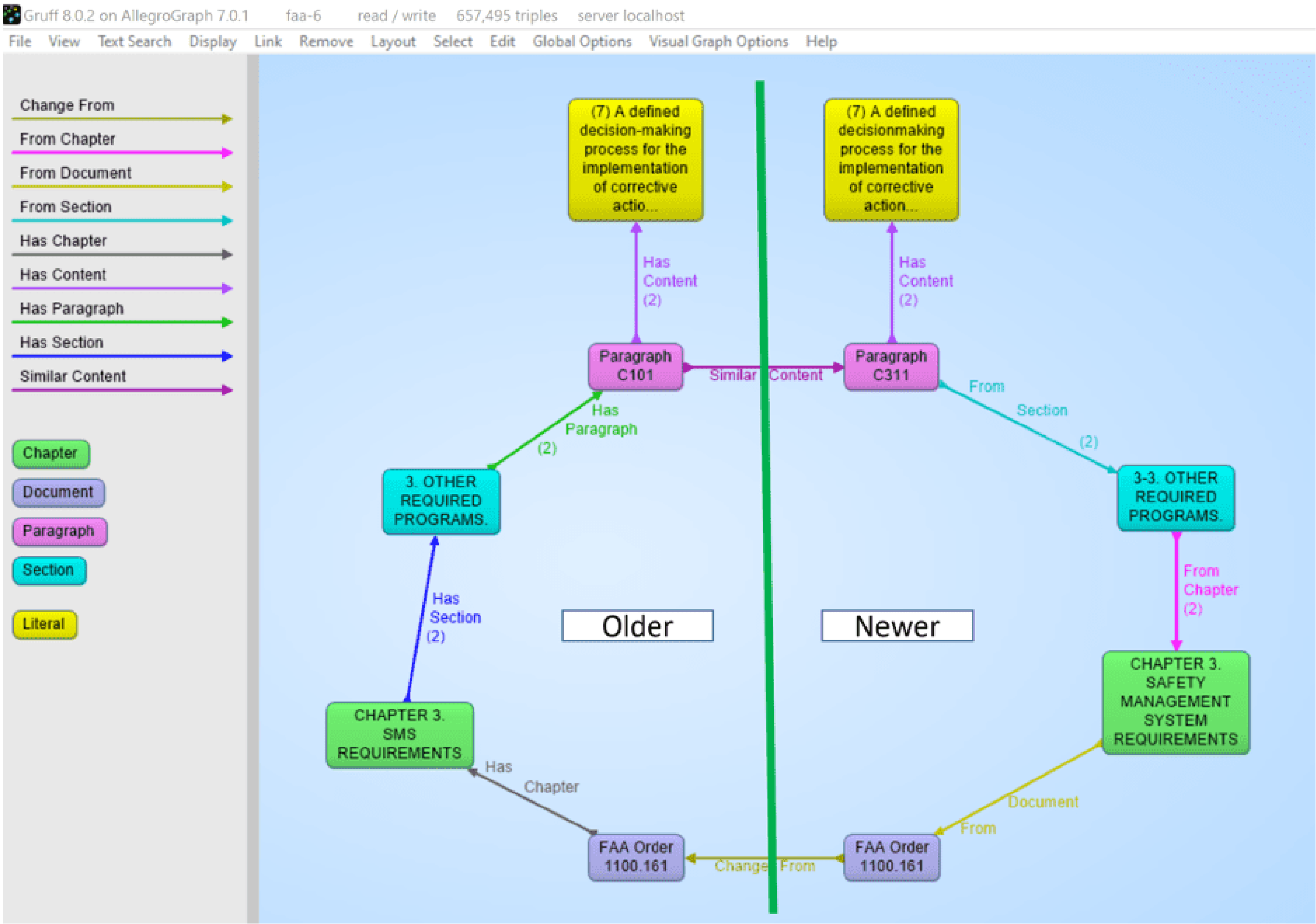
There are many many AI algorithms that can be applied in Document Knowledge Graphs. We provide best practices for topics like:

[a] Sentiment Analysis, using good/bad word lists or training data.
[b] Paragraph or Chapter similarity using statistical techniques like Gensim similarity or symbolic techniques where we just the overlap of recognized entities as a function of the size of a text.
[c] Query answering using word2vec or more advanced techniques like BERT
[d] Semantic search using the hierarchy in SKOS taxonomies.
[e] Summarization techniques for Abstractive or Extractive abstracts using Gensim or Spacy.

[9] Versioning and Document tracking

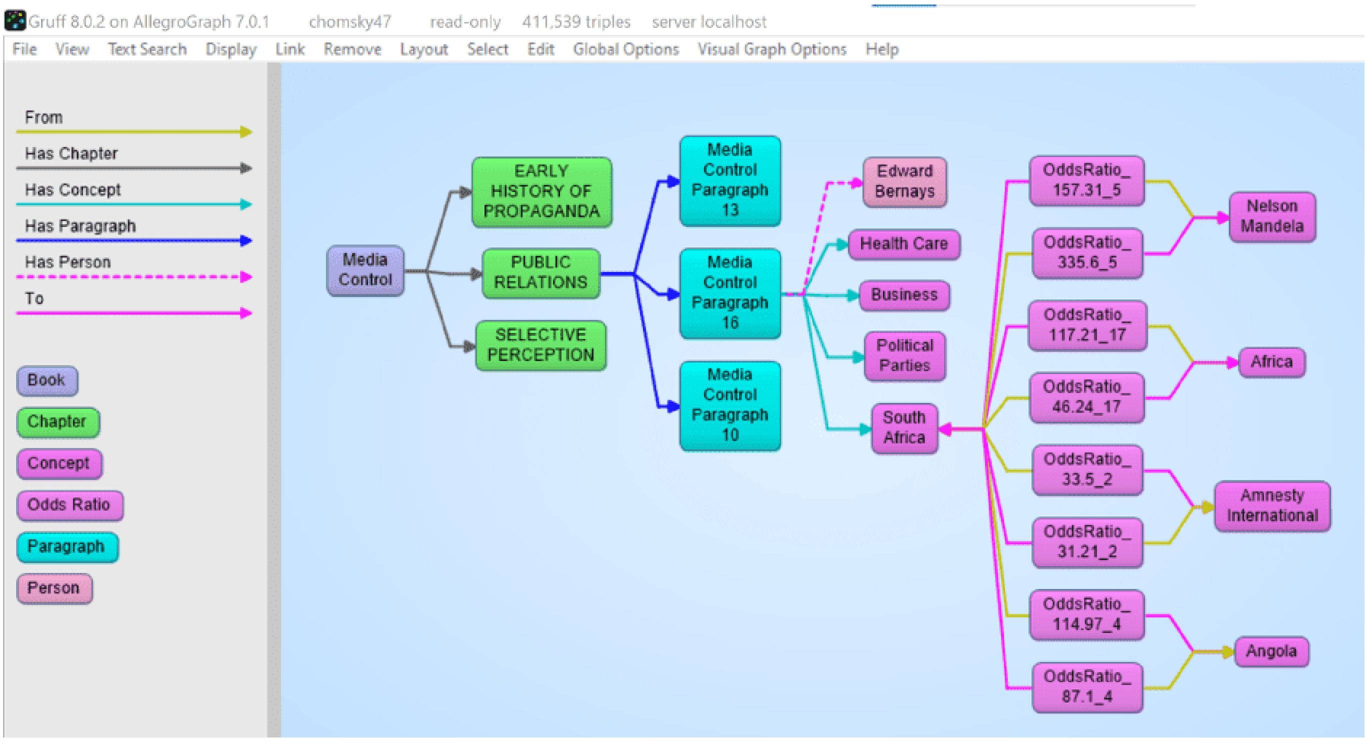
Several of our customers with Document Knowledge Graphs have noted the one constant in all of these KGs is that documents change over time. As part of our solution, we have created best practices where we deal with these changes. A crucial first step is to put each document in its own graph (i.e. the fourth element of every triple in the document tree is the document id itself). When we get a new version of a document the document ID changes but the new document will point back to the old version. We then compute which paragraphs stayed the same within a certain margin (there are always changes in whitespace) and we materialize what paragraphs disappeared in the new version and what new paragraphs appeared compared to the previous version. Part of the best practice is to put the old version of a document in a historical database that at all times can be federated with the ‘current’ set of documents.

Note that in the following picture we see the progression of a document. On the right hand side we have a newer version of a document 1100.161 with a chapter -> section -> paragraph -> contents where the content is almost the same as the one in the older version. But note that the newer one spells ‘decision making’ as one word whereas the older version said ‘decision-making’. Note that also the chapter titles and the section titles are almost the same but not entirely. Also, note that the new version has a back-pointer (changed-from) to the older version.



[10] Statistical Relationships

One important analytic one can do on documents is to look at the co-occurrence of terms. Although, given that certain words might occur more frequently in text, we have to correct the co-occurrence between words for the frequency of the two terms in a co-occurrence to get a better idea of the ‘surprisingness’ of a co-occurrence. The platform offers several techniques in Python and Lisp to compute these co-occurrences. Note that in the following picture we computed the odds ratios between recognized entities and so we see in the following gruff picture that if Noam Chomsky talks about South Africa then the chances are very high he will also talk about Nelson Mandela.


Gruff 8.0.2 on AllegroGraph 7.0.1
chomsky47
read-only
411,539 triples
server localhost
File
View
Text Search
Display
Link
Remove
Layout
Select
Edit
Global Options
Visual Graph Options
Help
From
Has Chapter
Has Concept
Has Paragraph
Has Person
To
Book
Chapter
Concept
Odds Ratio
Paragraph
Person
Media Control
EARLY HISTORY OF PROPAGANDA
PUBLIC RELATIONS
SELECTIVE PERCEPTION
Media Control Paragraph 13
Media Control Paragraph 16
Media Control Paragraph 10
Edward Bernays
Health Care
Business
Political Parties
South Africa
OddsRatio_157.31_5
OddsRatio_335.6_5
OddsRatio_117.21_17
OddsRatio_46.24_17
OddsRatio_33.5_2
OddsRatio_31.21_2
OddsRatio_114.97_4
OddsRatio_87.1_4
Nelson Mandela
Africa
Amnesty International
Angola