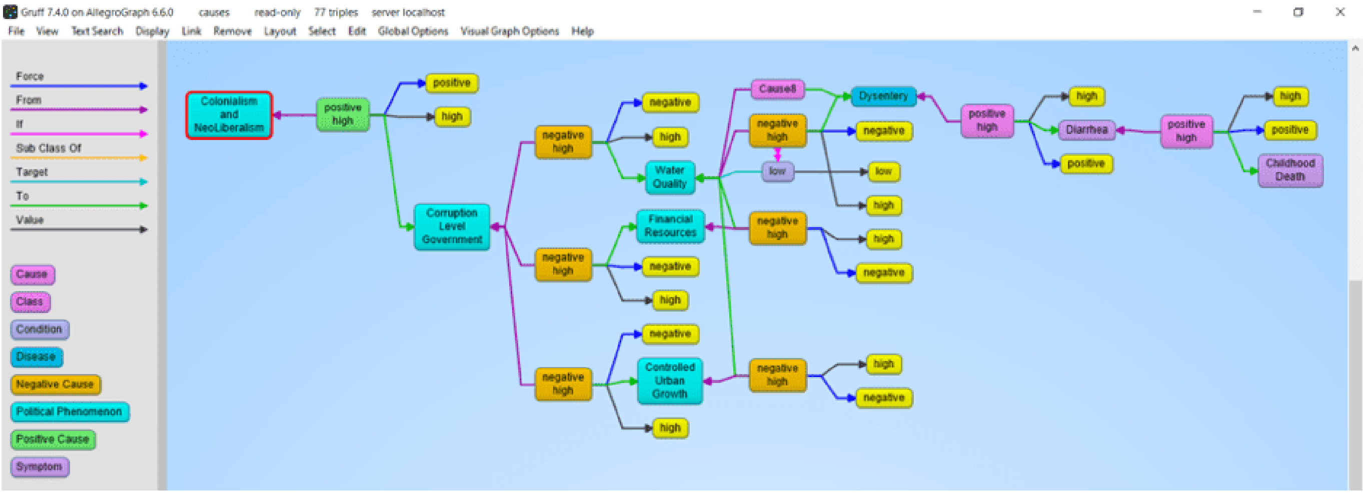
Document Knowledge Graphs with NLP and ML

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

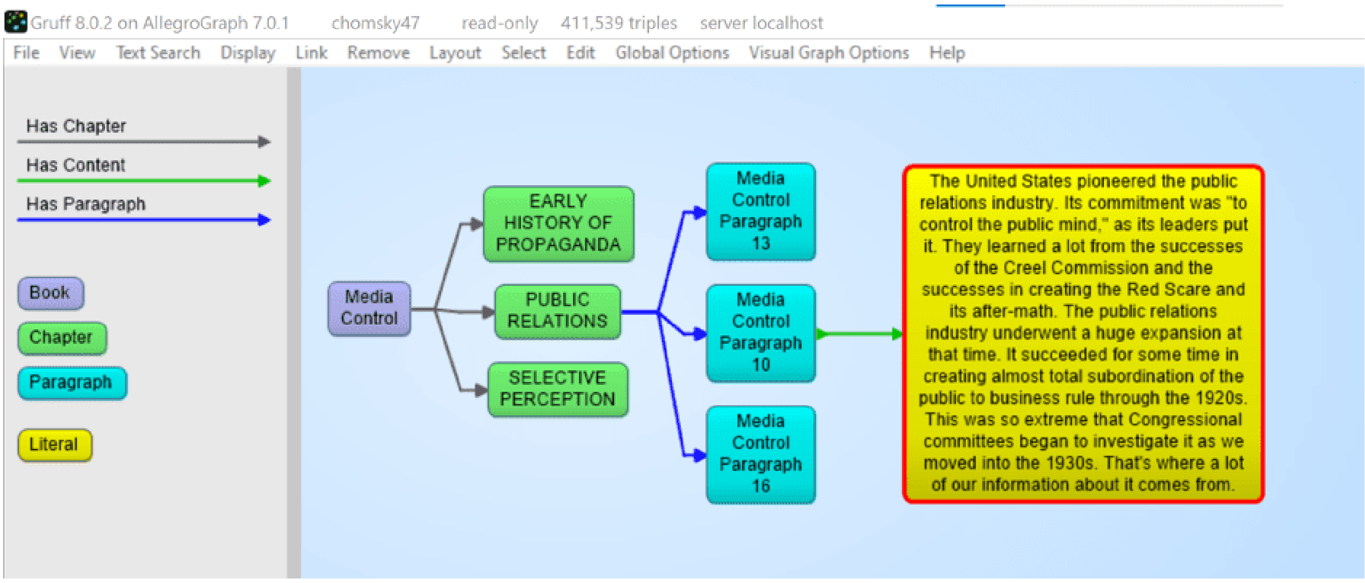
We then take a document and turn it into an intermediate XML using Apache Tika. Apache Tika supports more than 1000 document types and although Apache Tika is a fantastic tool, the output is still usually not clean enough to create a graph from, so we use Spacy rules to clean up the XML to make it as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

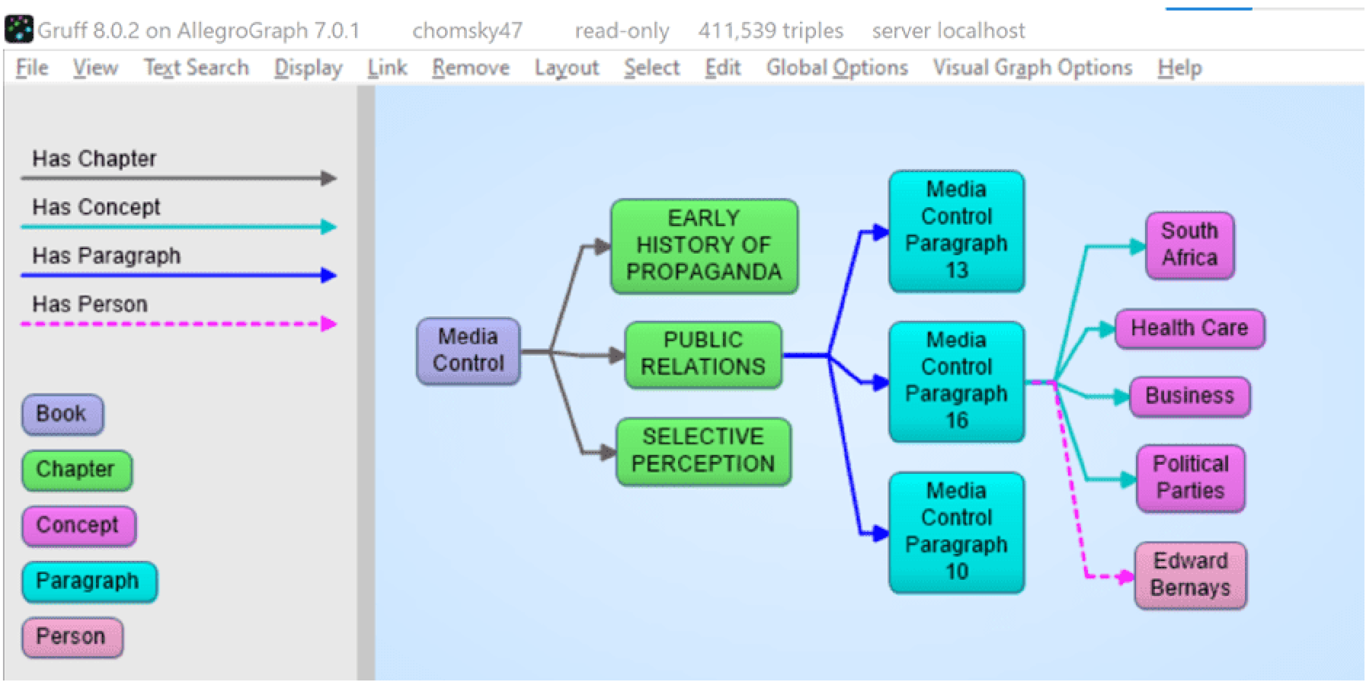
Our tools ingest the XML and metadata and transform that into a graph-based document tree. The document is the root and from that, it branches out into chapters, optionally sections, all the way down to paragraphs. The ultimate text content is in the paragraphs. In the following example we took the XML 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.

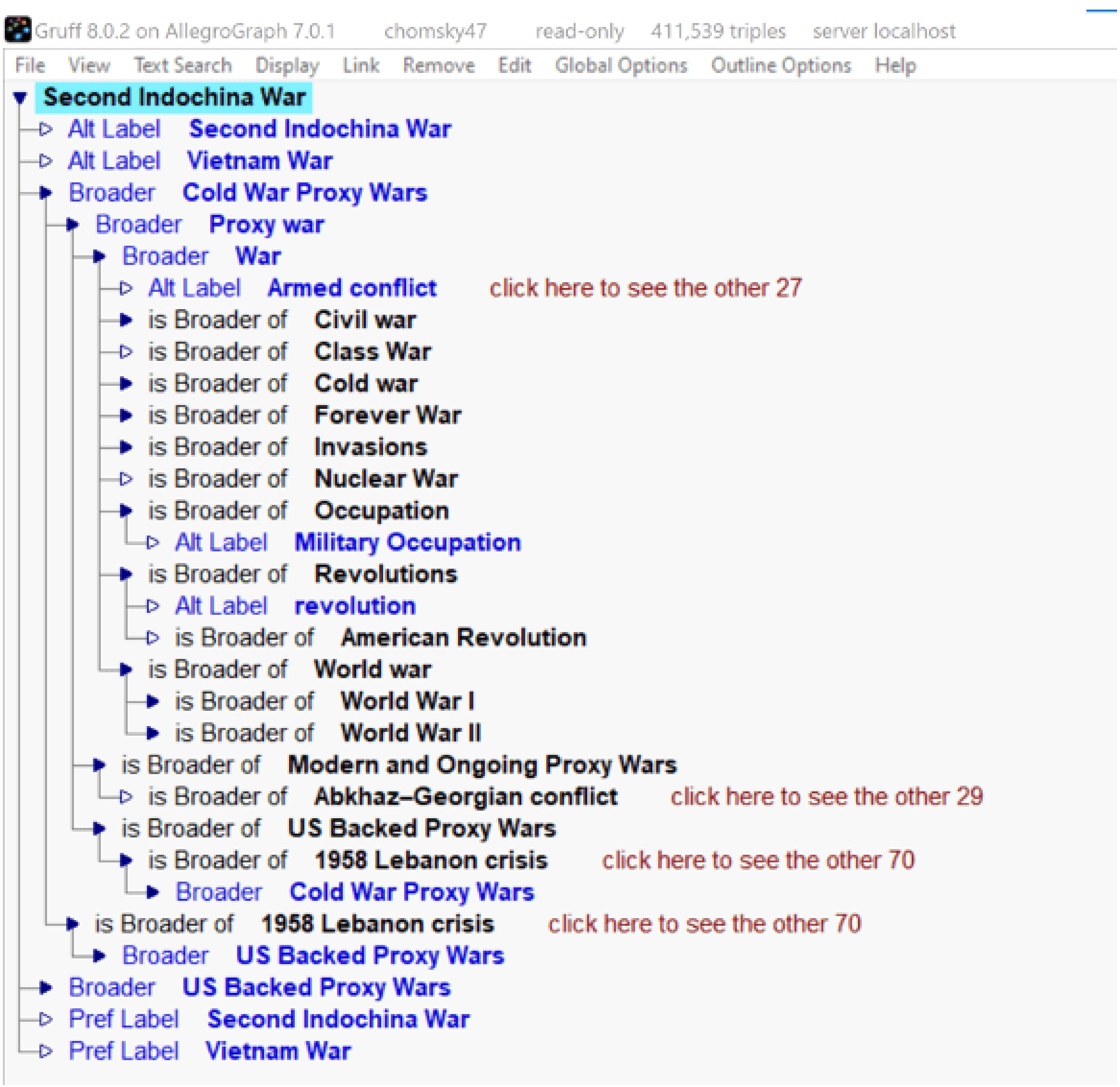


The United States pioneered the public relations industry. Its commitment was "to control the public mind," as its leaders put it. They learned a lot from the successes of the Creel Commission and the successes in creating the Red Scare and its after-math. The public relations industry underwent a huge 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.

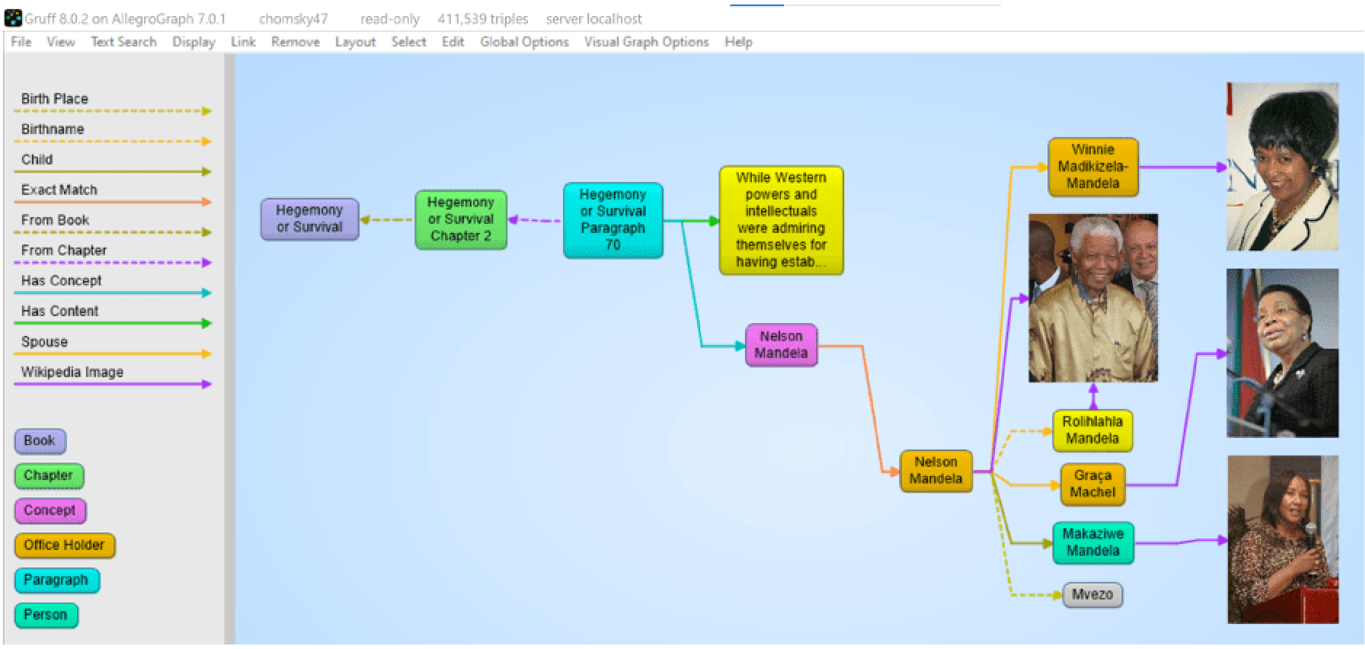

Gruff 8.0.2 on AllegroGraph 7.0.1 chomsky47 read-only 411,539 triples server localhost
File View Text Search Display Link Remove Layout Select Edit Global Options Visual Graph Options Help
Has Chapter
Has Concept
Has Paragraph
Has Person
Book
Chapter
Concept
Paragraph
Person
Media Control
EARLY HISTORY 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.



[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

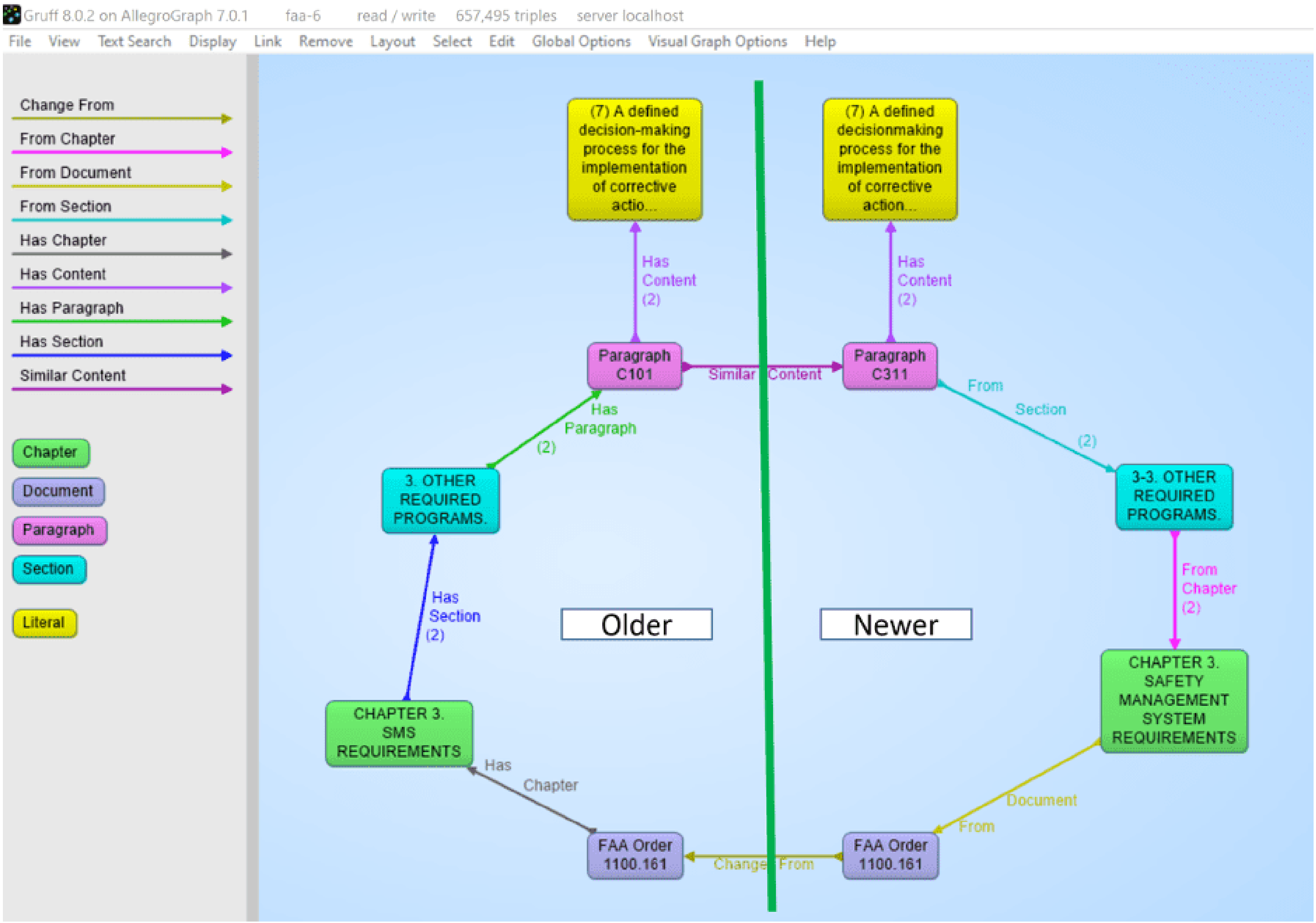
There are many many AI algorithms that can be applied in Document Knowledge Graphs. We provide best practices for topics like:

[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

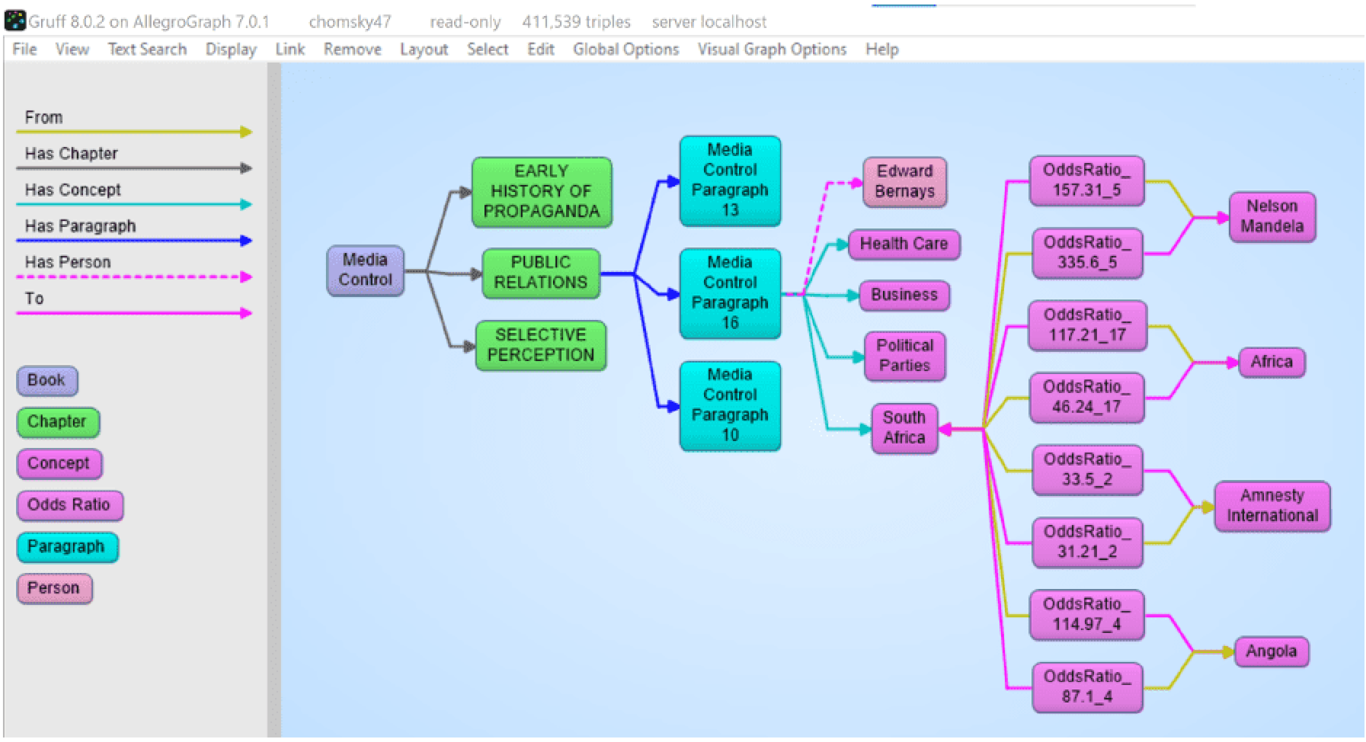
Several of our customers with Document Knowledge Graphs have noted the one constant in all of these KGs is that documents change over time. As part of our solution, we have created best practices where we deal with these changes. A crucial first step is to put each document in its own graph (i.e. the 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.



The Semantic Knowledge Graph: A Tribute

Dataversity – January 2019

Noam Chomsky, the philosopher, cognitive scientist, historian, social critic, and father of modern linguistics, has authored over 1,000 articles and 130 books. The 89-year-old intellectual also has written films and appeared in many documentaries. The substantial work he has done in linguistics and politics has earned him the title of “most cited living author.”

Now his work is the subject of the Noam Chomsky Knowledge Graph, the first Semantic Knowledge Graph for a public figure. “Doing a Semantic Project of all he has written or said is a fabulous tribute to a man who has made a big contribution to the study of language and its meaning,” says Fred Davis, Executive Director of the Chomsky Knowledge Graph project.

Dr. Jans Aasman was quoted:

“A good thing about Semantic Technology is that it’s easier to add new information than if you have a highly structured database. That’s because of the way things are stored in triples – where you have a subject, predicate, and object relationship—so you can bring in new information that instantly connects to other information,” says Dr. Jans Aasman, CEO of Franz.

Read the full article at Dataversity

Semantic Web and Semantic Technology Trends in 2019

Dataversity – January 2019

What to expect of Semantic Web and other Semantic Technologies in 2019? Quite a bit. DATAVERSITY engaged with leaders in the

space to get their thoughts on how Semantic Technologies will have an impact on multiple areas.

Dr. Jans Aasman, CEO of Franz Inc. was quoted several times in the article:

Among the semantic-driven AI ventures next year will be those that relate to the healthcare space, says Dr. Jans Aasman, CEO of Semantic Web technology company Franz, Inc:

“In the last two years some of the technologies were starting to get used in production,” he says. “In 2019 we will see a ramp-up of the number of AI applications that will help save lives by providing early warning signs for impending diseases. Some diseases will be predicted years in advance by using genetic patient data to understand future biological issues, like the likelihood of cancerous mutations – and start preventive therapies before the disease takes hold.”

If that’s not enough, how about digital immortality via AI Knowledge Graphs, where an interactive voice system will bring public figures in contact with anyone in the real world? “We’ll see the first examples of Digital Immortality in 2019 in the form of AI Digital Personas for public figures,” says Aasman, whose company is a partner in the Noam Chomsky Knowledge Graph:

“The combination of Artificial Intelligence and Semantic Knowledge Graphs will be used to transform the works of scientists, technologists, politicians, and scholars like Noam Chomsky into an interactive response system that uses the person’s actual voice to answer questions,” he comments.

“AI Digital Personas will dynamically link information from various sources – such as books, research papers, notes and media interviews – and turn the disparate information into a knowledge system that people can interact with digitally.”

These AI Digital Personas could also be used while the person is still alive to broaden the accessibility of their expertise.

On the point of the future of graph visualization apps, Aasman notes that:

"Most graph visualization applications show network diagrams in only two dimensions, but it is unnatural to manipulate graphs on a flat computer screen in 2D. Modern R virtual reality will add at least two dimensions to graph visualization, which will create a more natural way to manipulate complex graphs by incorporating more depth and temporal unfolding to understand information within a time perspective."

Read the full article at Dataversity.

Franz and Semantic Web Co. Partner to Create a Noam Chomsky Knowledge Graph

Press Release – September 10, 2018

First Semantic Knowledge Graph for a Public Figure will Semantically Link Books, Interviews, Movies, TV Programs and Writings from the Most Cited U.S. Scholar

OAKLAND, Calif. and VIENNA, Austria – Franz Inc., an early innovator in Artificial Intelligence (AI) and leading supplier of Semantic Graph Database technology, AllegroGraph, for Knowledge Graphs, and Semantic Web Company, developers of the PoolParty Semantic Suite and leading provider of Semantic AI solutions, today announced a partnership to develop the Noam Chomsky Knowledge Graph. This project is the first aimed at connecting all the works from a public figure and turning the linked information into a searchable and retrievable resource for the public.

The Noam Chomsky Knowledge Graph project will organize and semantically link the vast knowledge domain surrounding Noam Chomsky, the founder of modern linguistics, a founder of cognitive science, and a major figure in analytic philosophy as well as an American linguist, philosopher, historian and social critic. Chomsky is currently an Institute Professor Emeritus at the Massachusetts Institute of Technology (MIT) and laureate professor at the University of Arizona. He has received many awards including Distinguished Scientific Contribution Award of the American Psychological Association, the Kyoto Prize in Basic Sciences, the Helmholtz Medal, the Dorothy Eldridge Peacemaker Award, and the Ben Franklin Medal in Computer and Cognitive Science.

“Noam Chomsky is one of the most brilliant minds of our generation,” said Fred Davis. Executive Director of the Chomsky Knowledge Graph project, “His body of work is tremendously valuable to people across many disciplines. Our goal is to make Chomsky’s work searchable in the context of topics and concepts, readable in excerpts, and easily available to journalists, scientists, technologists, students, philosophers, and historians as well as the general public.”

The Noam Chomsky Knowledge Graph will link to over 1,000 articles and over 100 books that Chomsky has authored about linguistics, mass media, politics and war. Hundreds of Chomsky’s media interviews, which aired on television, print and online will be part of the Knowledge Graph as well as more than a dozen Chomsky movies including: Is the Man who is Tall Happy?, Manufacturing Consent, Programming the Nation? Hijacking Catastrophe: 911 Fear and the Selling of American Empire. The content will be made available by searching the Knowledge Graph for specific titles, related topics and concepts.

Since the project is based on the latest and most advanced technologies, the data will be also available as machine-readable data (Linked Data) in order to be fed into smart applications, intelligent chatbots, and question/ answering machines – as well as other AI and data systems.

The Internet Archive, the world's largest digital lending library, will host Noam Chomsky's books, movies, and other content – enabling public access to his works and marking the first integration between the Internet Archive and a public Knowledge Graph.

"We are thrilled to be working on this momentous project," said Dr. Jans Aasman, CEO of Franz Inc. "Noam Chomsky is the ideal person to fulfill the vision of a Public Figure Knowledge Graph. We are looking forward to collaborating with the Semantic Web Company and Fred Davis on this exciting project."

"Knowledge Graphs are becoming increasingly important for addressing various data management challenges in industries such as financial services, life sciences, healthcare or energy," said Andreas Blumauer, CEO and founder of Semantic Web Company. "The application of Knowledge Graphs to public figures, such as Noam Chomsky, will offer a unique opportunity to link concepts and ideas to form new ideas and possible solutions."

About Knowledge Graphs

A Knowledge Graph represents a knowledge domain and connects things of different types in a systematic way. Knowledge

Graphs encode knowledge arranged in a network of nodes and links rather than tables of rows and columns. People and machines can benefit from Knowledge Graphs by dynamically growing a semantic network of facts about things and use it for data integration, knowledge discovery, and in-depth analyses.

Gartner recently identified Knowledge Graphs as a key new technology in both their Hype Cycle for Artificial Intelligence and Hype Cycle for Emerging Technologies. Gartner’s Hype Cycle for Artificial Intelligence, 2018 states, “The rising role of content and context for delivering insights with AI technologies, as well as recent knowledge graph offerings for AI applications have pulled knowledge graphs to the surface.”

Knowledge Graphs are the Foundation for Artificial Intelligence

The foundation for AI lies in the facets Knowledge Graphs and semantic technology provided by Franz and Semantic Web Company. The Franz AllegroGraph Semantic Graph database provides the core technology environment to enrich and contextualized the understanding of data. The ability to rapidly integrate new knowledge is the crux of the Knowledge Graph and depends entirely on semantic technologies.

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 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. For more information, visit www.franz.com.

About Semantic Web Company

Semantic Web Company is the leading provider of graph-based metadata, search and analytic solutions. The company is the vendor of PoolParty Semantic Suite, one of the most renowned semantic software platforms on the global market. Among many other customers, The World Bank, AT&T, Deutsche Telekom, and Pearson benefit from linking structured and unstructured data. In 2018, the Semantic Web Company has been named to KMWorld's "100 companies that matter in Knowledge Management." For more information about PoolParty Semantic Suite, please visit https://ww.poolparty.biz

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

Allegro Knowledge Graph News

Franz periodically distributes newsletters to its Semantic Technologies, and Common Lisp based Enterprise Development Tools mailing lists, providing information on related upcoming events and new software product developments.

Read our latest AllegroGraph newsletter.

Previous issues are listed in the Newsletter Archive.

AllegroGraph News

Franz periodically distributes newsletters to its Semantic Technologies, and Common Lisp based Enterprise Development Tools mailing lists, providing information on related upcoming events and new software product developments.