Knowledge Graphs

Concept, mogelijkheden en aandachtspunten

Christophe Debruyne
Smals Research
30 maart 2021

Smals
ICT for society

Smals Research 2021

Innovation with new technologies

Consultancy & expertise

Internal & external knowledge transfer

Support for going live

HA Software Architectures

Anomalies & Transaction Management

Conversational Platform

On-premise Cloud

GIS for Analytics

Natural Language Generation

Knowledge Graphs

2021

Knowledge Retention

Next Best Action

AI Deployment

European Blockchain Infrastructure

Privacy Enhancing Technologies

Passwordless Authentication

"Proof of Life"

Doel van de webinar

Het beantwoorden van de volgende vragen:

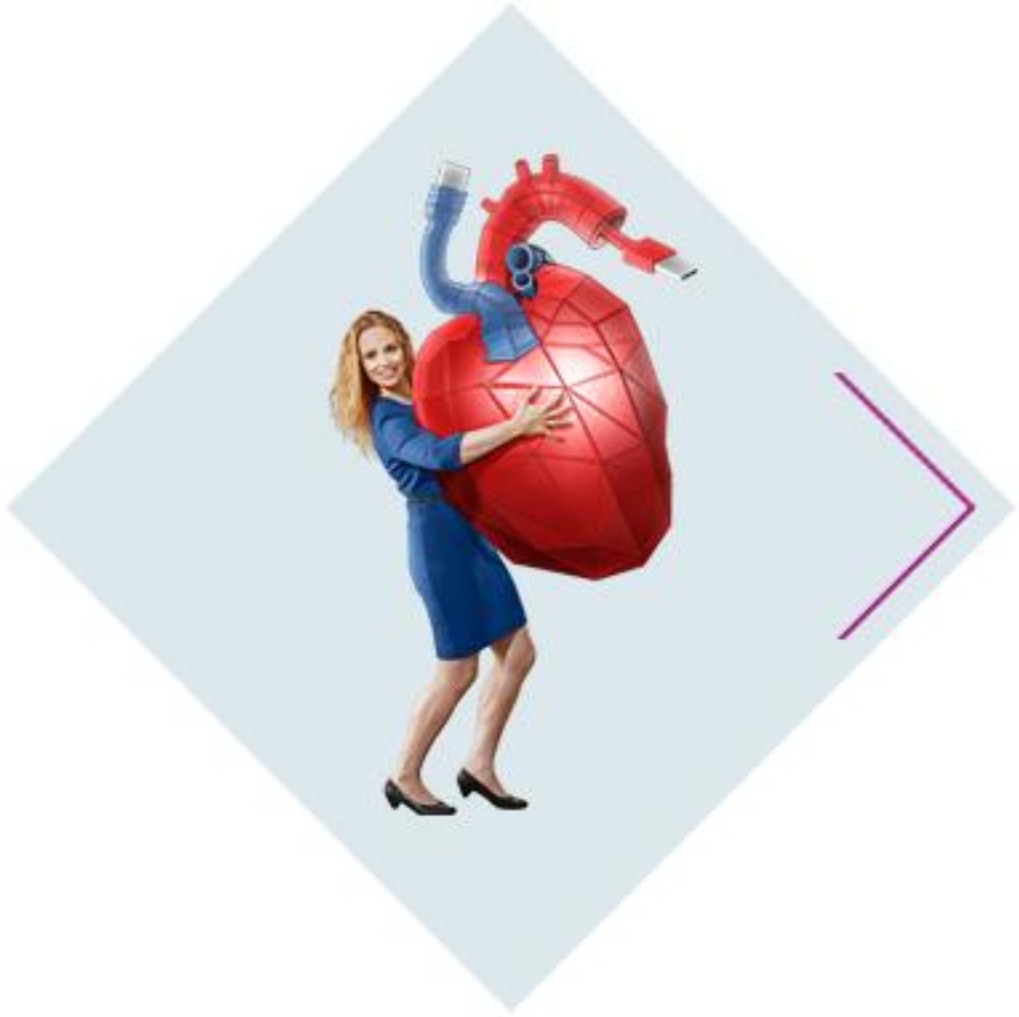
- Wat is een knowledge graph (KG)?
- Waar en voor welke problemen kan men KG's toepassen?
- Wat komt er kijken bij het bouwen en onderhouden van een KG?

Het kort toelichten van de volgende vragen:

- Hoe verhouden KG's zich tegenover andere initiatieven?
- Hoe verhouden KG's zich tegenover andere AI technieken?


https://www.pexels.com/photo/flying-hot-air-balloon-above-snow-covered-mountain-1740103/

Smals
ICT for society



Het concept

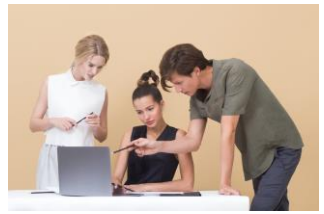
Wat is een
Knowledge Graph?

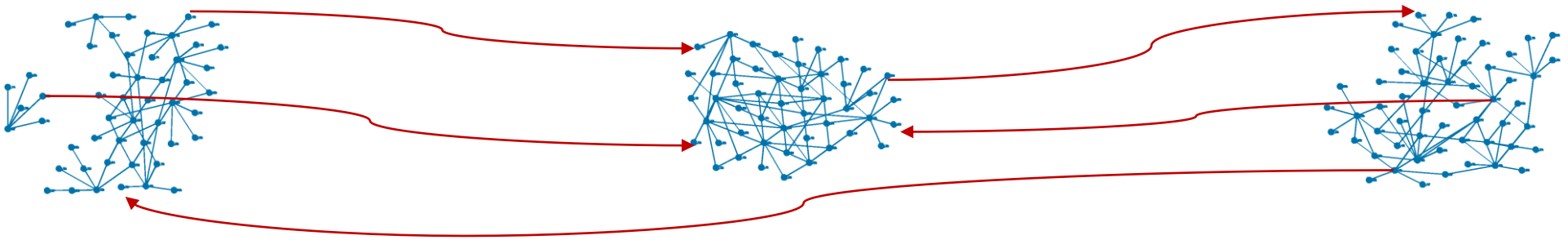
Probleem

Een schat van kennis is verdoken, is niet geïntegreerd, en is dus niet maximaal benut

De kennis is niet expliciet, niet gestructureerd, en bijgevolg moeilijk te delen of te gebruiken

Ongestructureerde data	Gestructureerde data	Kennis van experten





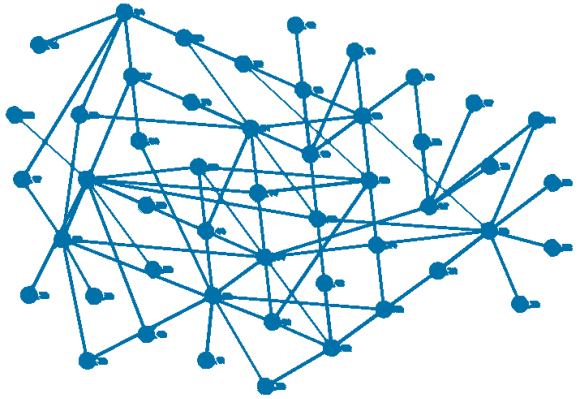
… over diensten, processen, departementen, organisaties, etc. heen

Een knowledge graph structureert die kennis en maakt de verbanden expliciet

Smals ICT for society

Wat is een knowledge graph?

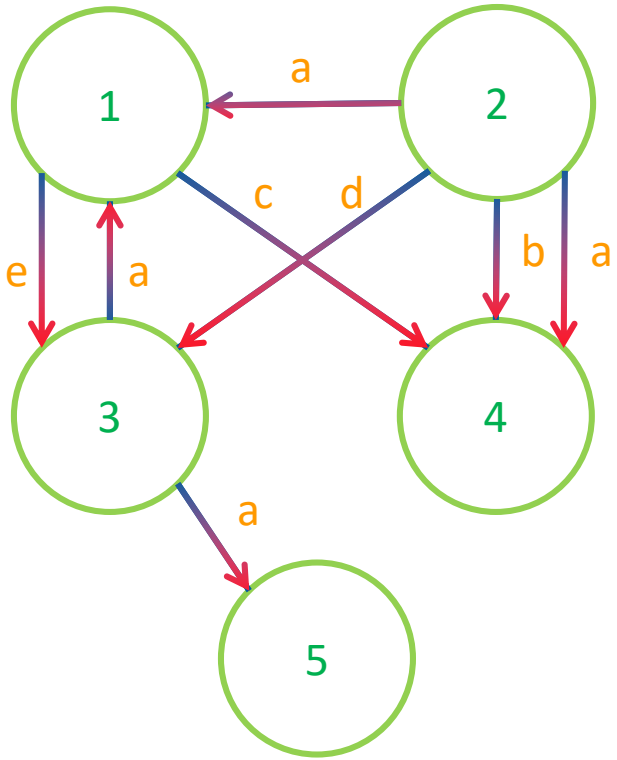
Een KG bestaat uit een graph die aan bepaalde voorwaarden voldoet



https://www.pexels.com/photo/banking-business-checklist-commerce-416322/

Smals
ICT for society

Wat is de wiskundige notie van een graaf?



Een graaf is een belangrijk concept in de wiskunde en de informatica

Een graaf bestaat uit:

Verzameling nodes (vertices, knopen), en een

Verzameling edges (zijden) tussen die nodes

In een gerichte graaf hebben edges een richting

Nodes krijgen doorgaans unieke labels

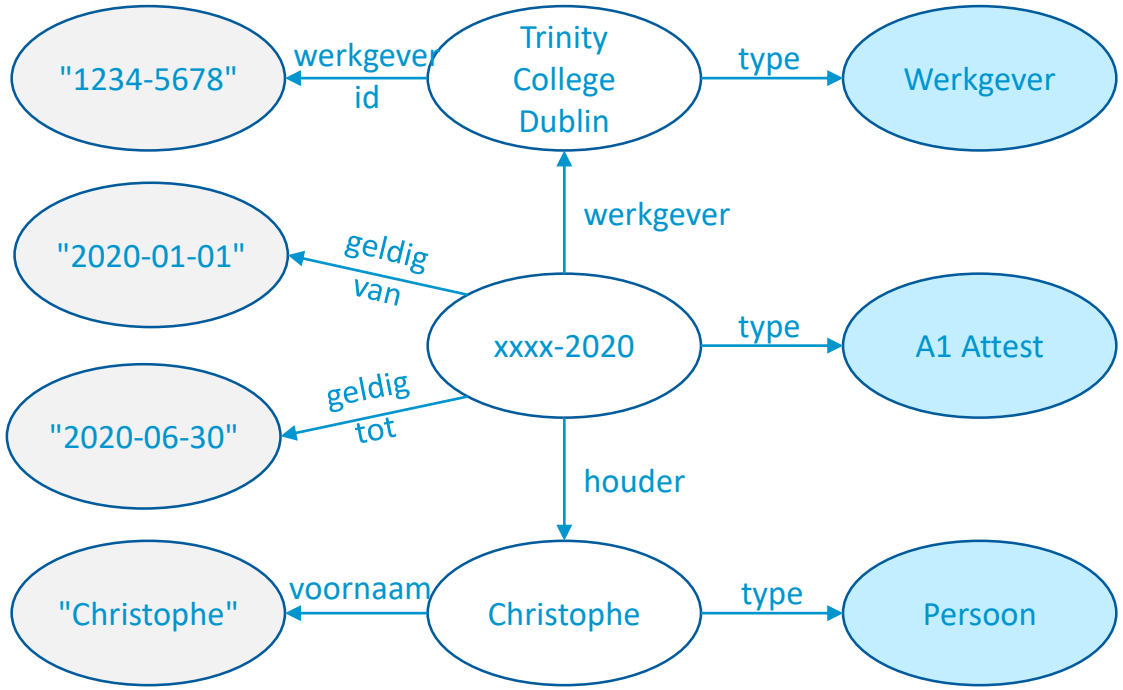
Edges kunnen ook gelabeld worden

Hoe gebruiken we een graaf om kennis voor te stellen?

Smals ICT for society

De graph van een knowledge graph

Entiteiten	Stellen "dingen" in een domein
Relaties →	Relaties tussen "dingen"
Typen	Entiteiten voor categorieën Geven context en betekenis
Attributen	Entiteiten voor tekenreeksen; data, nummers, …



Smals ICT for society

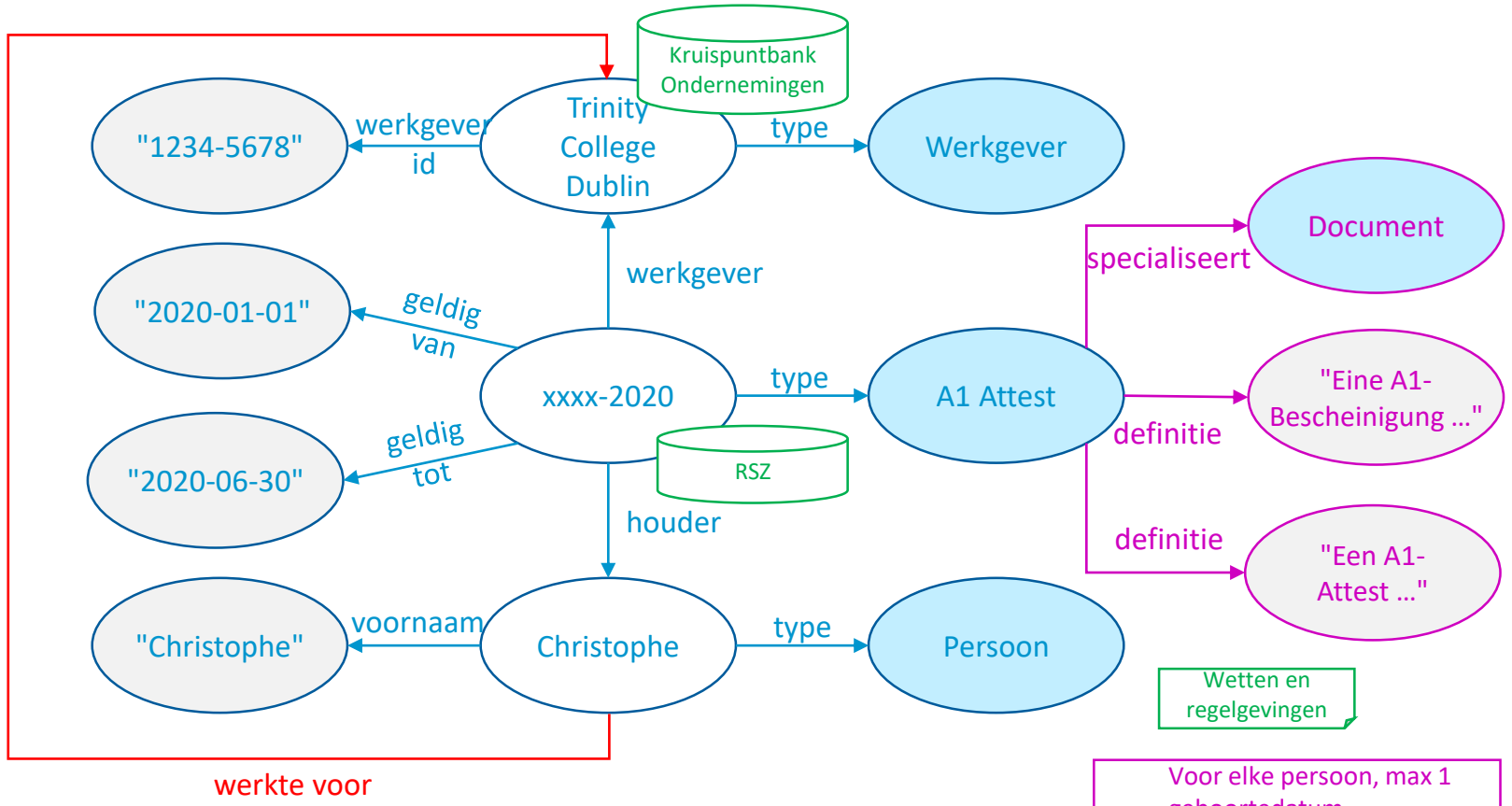
De voorwaarden van een knowledge graph

Entiteiten	Stellen "dingen" in een domein
Relaties	Relaties tussen "dingen"
Typen	Entiteiten voor categorieën Geven context en betekenis
Attributen	Entiteiten voor tekenreeksen; data, nummers, …

1. Typen en relaties worden in een schema gedocumenteerd (definities, eigenschappen)

2. Integratie van informatie van verschillende domeinen, organisaties, departementen, en zelfs uit verschillende bronnen

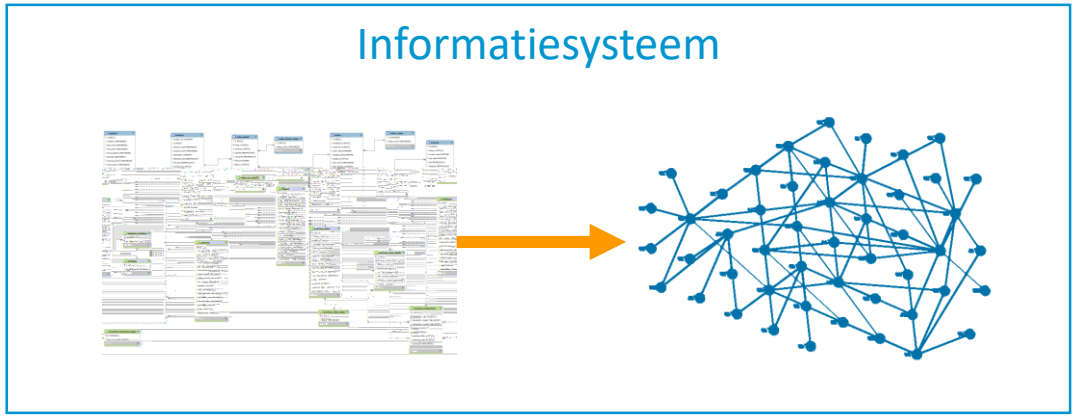
3. Ondersteuning om impliciete relaties, inzichten, kennis,… af te leiden



De voorwaarden van een knowledge graph

Graph technologieën zijn nuttig

E.g., efficiëntie in bepaalde use cases



Dus: niet elke graph is een knowledge graph

De voorwaarden laten ons toe niet alleen om graphs te onderscheiden, maar ook in te schatten wanneer een knowledge graph toepasselijk is

Bron ERD diagram: https://github.com/dalers/mywind/blob/master/northwind-erd.png

Oude wijn in nieuwe zakken

Klinken de definitie en beschrijving van een KG bekend in de oren?

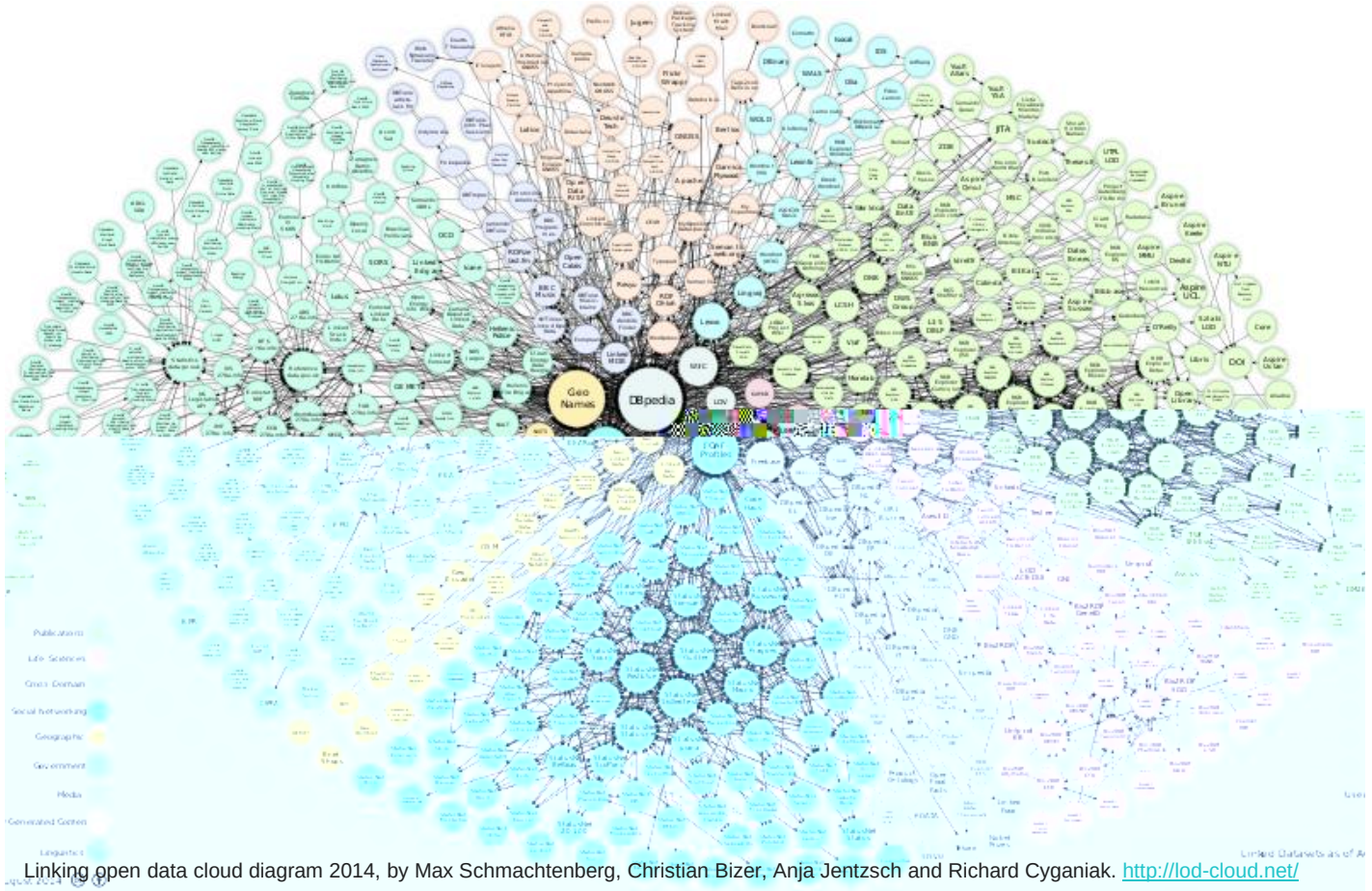
- Semantic Web
- RDF en ontologieën
- Knowledge bases
- Linked Data
- …

Maar waarom de "hernieuwde" interesse?

Vier redenen (persoonlijke opinie)

Hernieuwde interesse

Eind jaren '90	Semantic Web
Vanaf Midden 2000	Linked Data



Linking open data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

Hernieuwde interesse

Eind jaren '90 Semantic Web

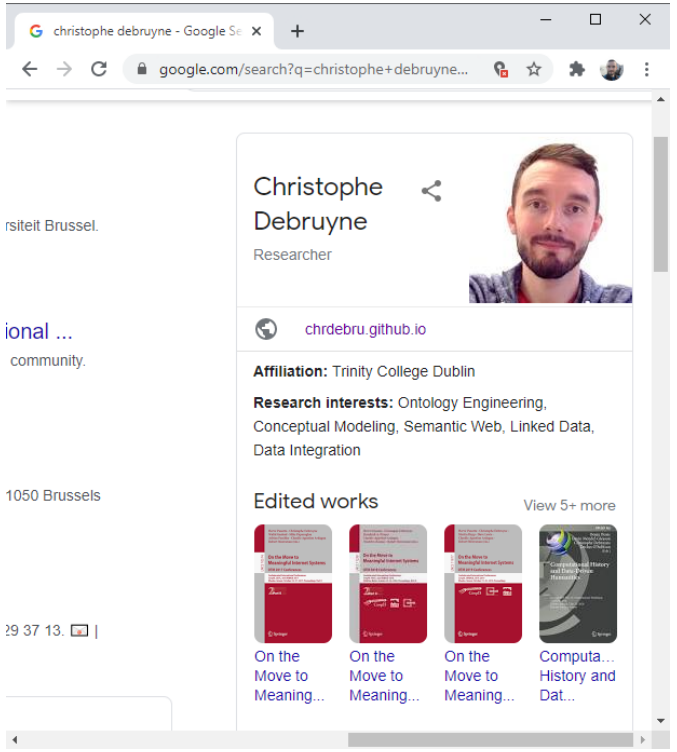
Vanaf Midden 2000 Linked Data

Reden 1: "Rebranding"

Google maakte de term "Knowledge Graph" in 2012 populair

Hoewel de "knowledge graphs" sinds de jaren 70 gebruikt

Vanaf 2012 Linked Data ~~en~~ ⊂ KG's



KG's zijn niet gelijk aan Linked Data, doch worden Linked Data technologieën en principes vaak voor KG's gebruikt

Ruimte voor pragmatische keuzes

Smals ICT for society

Hernieuwde interesse

Reden 2: De (onderliggende) technologieën werden toegankelijker

Twee voorbeelden

1. Schema.org (gedreven door Google, Microsoft, Yahoo!, en Yandex)
   Toegankelijker voor (Web) ontwikkelaars
   → Incentive voor bedrijven
   → Kennisinstellingen bieden vakken aan

2. JSON-LD ("Linked Data in JSON formaat")
   Ontwikkelaars helpen bij het produceren en gebruiken van Linked Data

Maar… de "kunst" van het modelleren gaat "verloren" (een andere discussie)

Hernieuwde interesse

Reden 3: Onderzoek naar graph datamodellen en formalismen ligt niet stil

De relaties tussen verschillende soorten graph datamodellen

Schematalen, validatietalen, …

Graph database technologieën (NoSQL), virtualisatie, …

Hernieuwde interesse

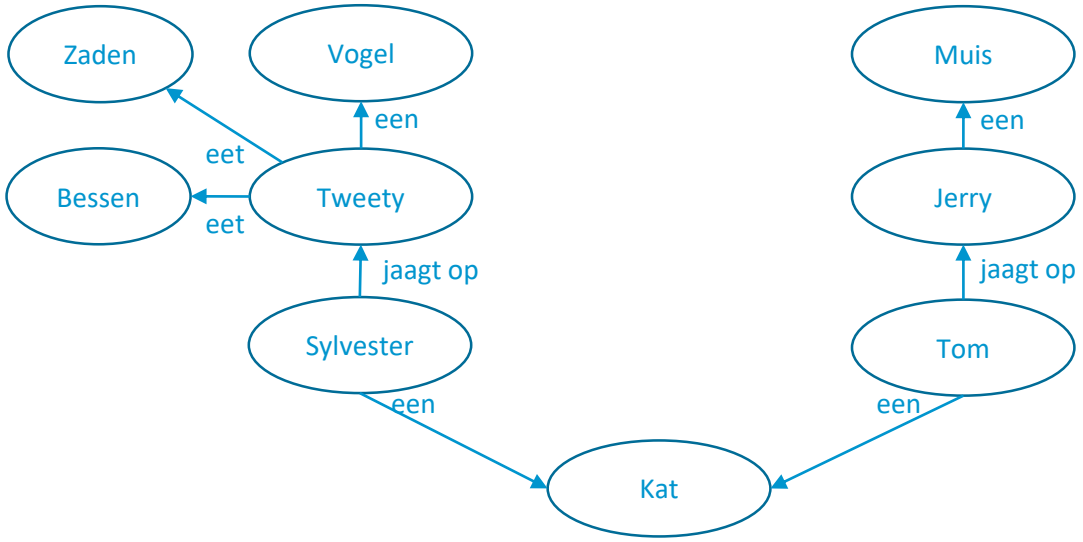
Reden 4: Recente ontwikkelingen in IT

Big data processing en opslag in de cloud

Noot: niet alle KG's hebben nood aan Cloud oplossingen

Machine Learning en Deep Learning met KG's én schema's

M/DL op de graph, en M/DL met de graph



Een graph biedt een meer flexibele manier om gegevens voor te stellen.

Graph Neural Networks (GNN) codeert de informatie van een node's "buren" voor verschillende taken: labels voorspellen, nodes voorspellen, relaties voorspellen,…

Hier ziet men dat katten jagen, dus bij het introduceren van een nieuwe kat (e.g., Azrael), kan men voorspellen dat Azrael op *iets* zal jagen.

Smals ICT for society

Hernieuwde interesse

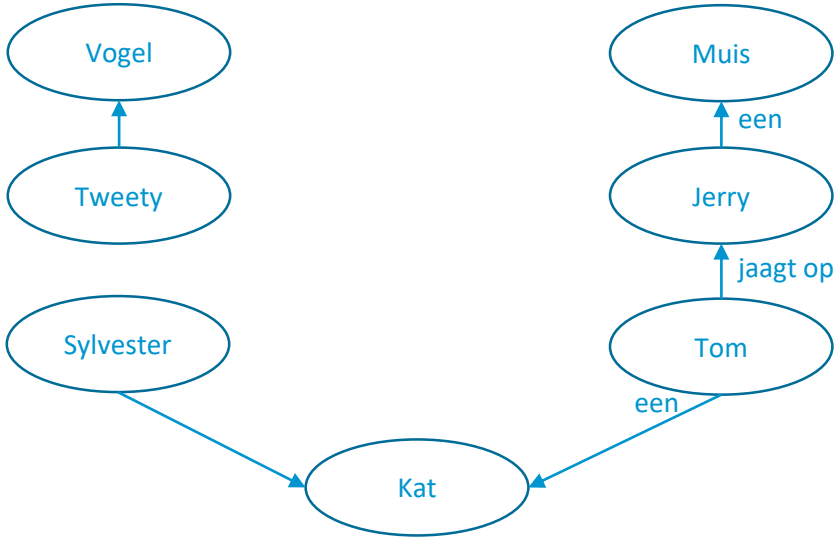
Reden 4: Recente ontwikkelingen in IT

Big data processing en opslag in de cloud

Noot: niet alle KG's hebben nood aan Cloud oplossingen

Machine Learning en Deep Learning met KG's én schema's

M/DL op de graph, en M/DL met de graph



Smals ICT for society

Enterprise Knowledge Graphs

Hoe verhouden KGs in een bedrijf/organisatie zich tegenover open KG's en Linked Data?

Features	Open KGs / Linked Data	Enterprise KGs
Data source	Distributed	Usually centralized
Openness	Open to public	Private
Size of the data	(Usually) huge	(Usually) big
Data *Acquisition*	Usually harder	Usually easier
Quality of the data	Low	High(er)
Schema language	Simple	Likely to be more expressive
Knowledge	Generic or Domain Specific	Domain-specific

Gebaseerd op Pan et al., 2017

Smals ICT for society

Toepassingen

Waar en waarvoor worden knowledge graphs gebruikt?

+

Concrete voorbeelden

Waarom Knowledge Graphs?

Semantic Search & Information Retrieval

Recommender Systems

Kennisretentie

Chatbots en Virtual Assistants

Business en Financial Analytics

Meertalige diensten

Operational Research (logistics)

Drug Discovery

…

Welke soort problemen?

Een oplossing omvat:

- Bewaren en toegankelijk maken van kennis
- Afleiden van kennis en inzichten

Uit informatie van:

- Verschillende domeinen
- Verschillende bronnen
- Niet expliciet opgeslagen
- …

Problemen en applicaties hoeven niet groot te zijn!

Waarom Knowledge Graphs?

Semantic Search & Information Retrieval

Recommender Systems

Kennisretentie

Chatbots en Virtual Assistants

Business en Financial Analytics

Meertalige diensten

Operational Research (logistics)

Drug Discovery

…

Intelligente bevragingen via de KG door context te achterhalen. KG bevat ook synoniemen en exploiteert eigenschappen van relaties.

"Wie zit er op de bank?"

Smals ICT for society

Waarom Knowledge Graphs?

Semantic Search & Information Retrieval

Recommender Systems

Kennisretentie

Chatbots en Virtual Assistants

Business en Financial Analytics

Meertalige diensten

Operational Research (logistics)

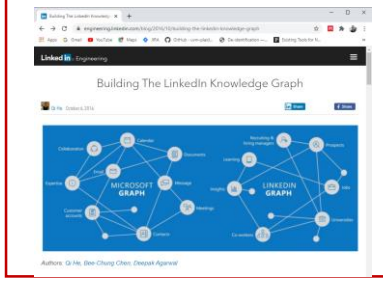
Drug Discovery

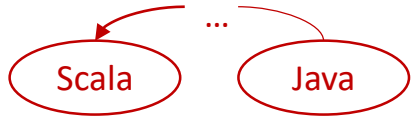
…

Intelligente bevragingen via de KG door context te achterhalen. KG bevat ook synoniemen en exploiteert eigenschappen van relaties.

"Wie zit er op de bank?"

Aanbevelingen van bv documenten op basis van gelijkaardige activiteiten. KG integreert doorgaans meerdere bronnen in dit scenario.





https://www.linkedin.com/

Smals ICT for society

Waarom Knowledge Graphs?

Semantic Search & Information Retrieval

Recommender Systems

Kennisretentie

Chatbots en Virtual Assistants

Business en Financial Analytics

Meertalige diensten

Operational Research (logistics)

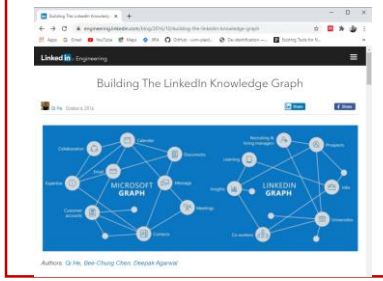
Drug Discovery

…

Intelligente bevragingen via de KG door context te achterhalen. KG bevat ook synoniemen en exploiteert eigenschappen van relaties.

"Wie zit er op de bank?"

Aanbevelingen van bv documenten op basis van gelijkaardige activiteiten. KG integreert doorgaans meerdere bronnen in dit scenario.



https://www.linkedin.com/

KG's als onderdeel van een bredere strategie voor het kwantificeren, bewaren, en toegankelijk maken van kennis en expertise (turnover, pensioen, etc.).

Smals ICT for society

Voorbeelden van grote projecten

BASF
The Chemical Company

Integreren van > 200 miljoen documenten, data, data van eigen experimenten, etc. om R&D te drijven. E.g., de functie van genen voorspellen.

NETFLIX

Netflix's Content Knowledge Graph wordt gebruikt voor, onder andere, aanbevelingen (op basis van inhoud, gebruikersactiviteit, en kennis).

KG projecten hoeven niet groot te zijn.

En als men een KG project aanvangt: **klein beginnen**.

AstraZeneca

Integreren van documenten voor aanbevelingen, dit om het proces van drug discovery te versnellen.

SIEMENS
Ingenuity for life

Een "industriële" KG voor "Siemens Domain Knowledge". Data uit verschillende silo's ontsluiten en verrijken. Doel? Data toegankelijker maken en processen optimaliseren.

https://medium.com/dataseries/how-knowledge-graphs-will-transform-data-management-and-business-2b0aad9b5342
https://towardsdatascience.com/movie-recommendations-powered-by-knowledge-graphs-and-neo4j-33603a212ad0
https://www.ns-healthcare.com/analysis/astrazeneca-drug-discovery/
https://indico.cern.ch/event/669648/contributions/2838194/attachments/1581790/2499984/CERN_Open_Lab_Technical_Workshop_-_SIEMENS_AG_-_FISHKIN_-_11-01-2018.pdf

Smals
ICT for society

RTÉ De Ierse nationale zender

Journalisten, onderzoekers (voor programma's), etc. hadden het moeilijk archieven te doorzoeken

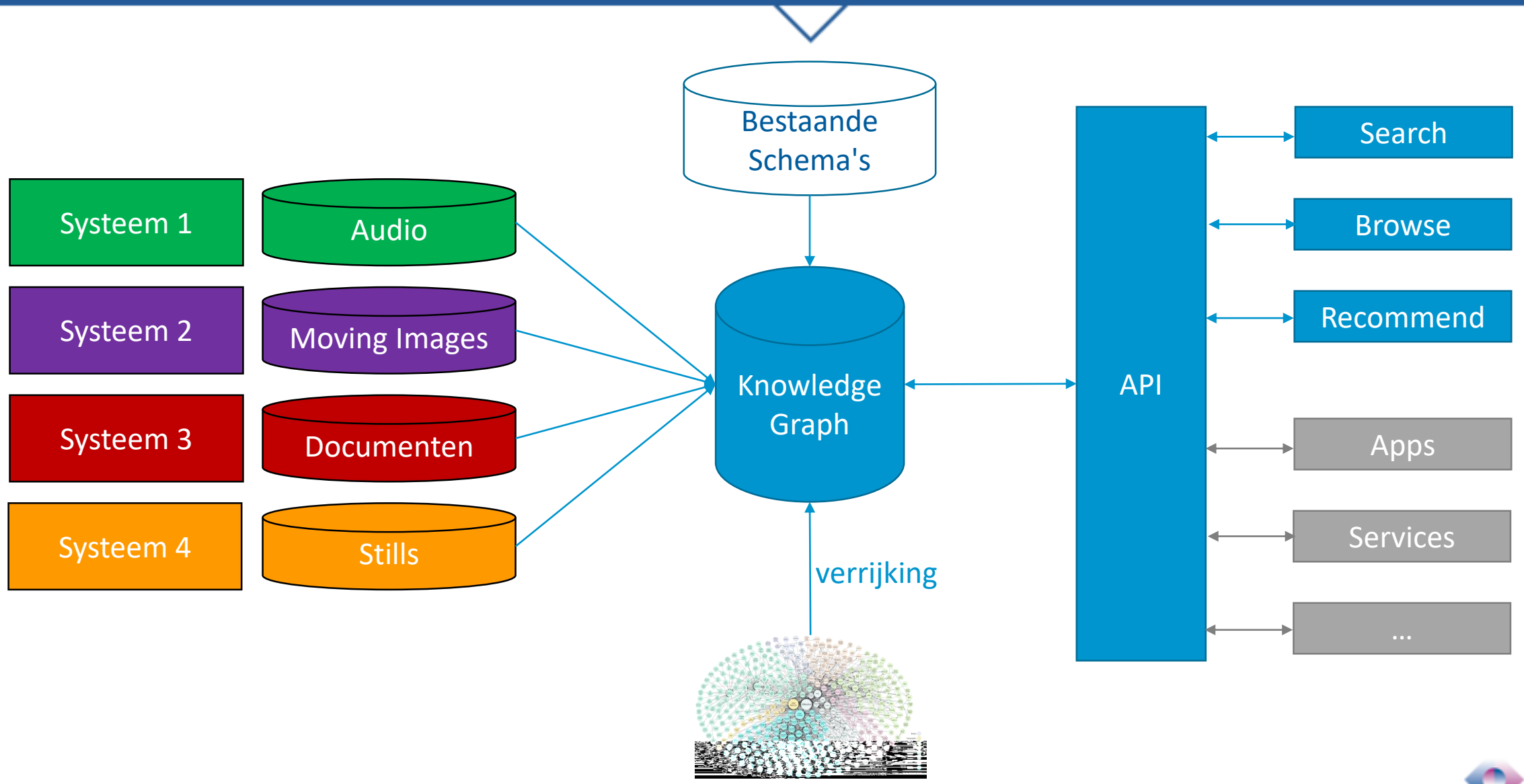
Uitdagingen:

- Vier verschillende archieven, systemen, en procedures
- Data kwaliteit
- Focus op metadata, dus…
- Zeer tot weinig context

Vereisten:

- Een centraal punt voor opzoekingen op basis van bestaande data
- Meer context
- Bestaande systemen niet aanraken
- Derden applicaties laten ontwikkelen

RTÉ De Ierse nationale zender

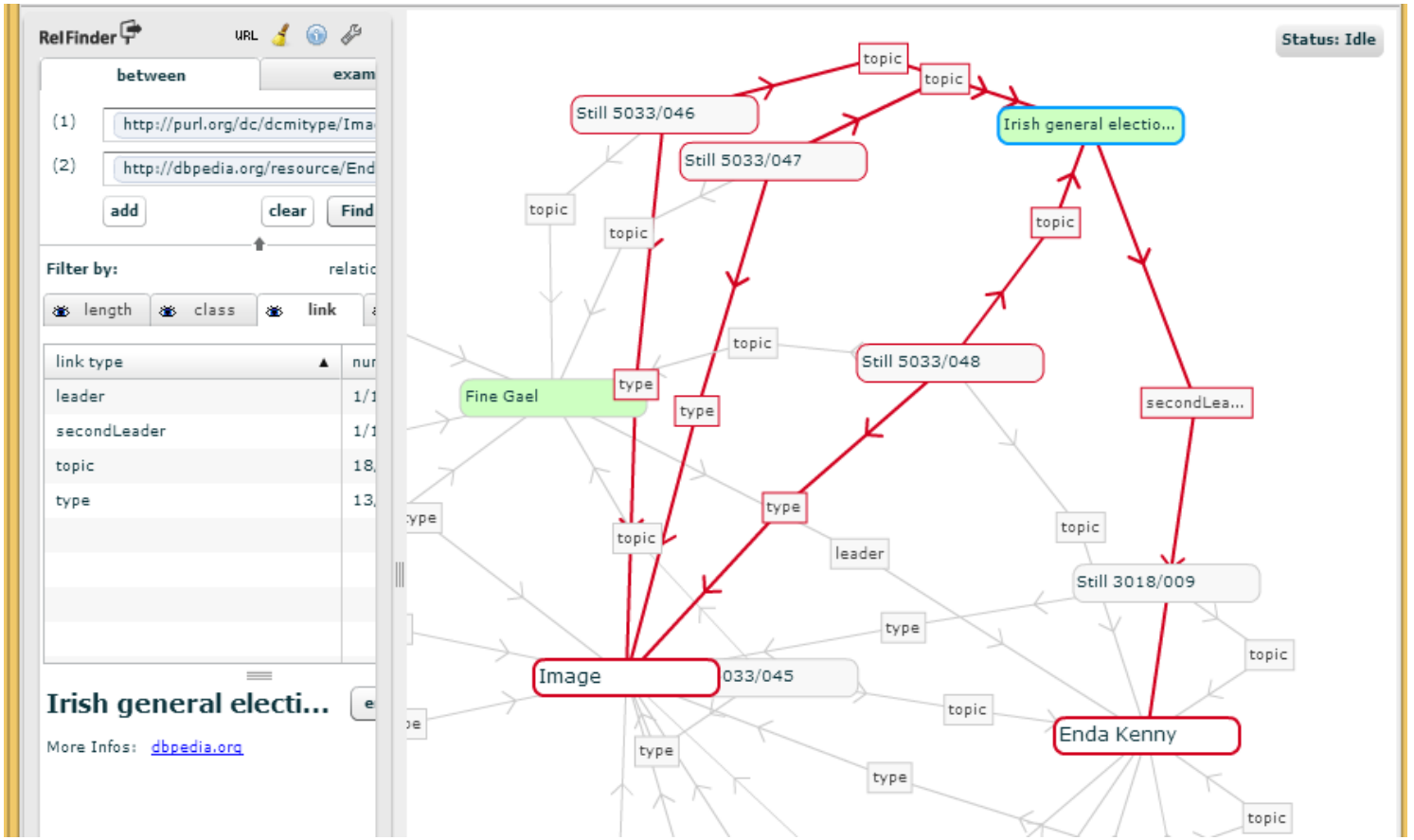


Smals ICT for society

RTÉ De Ierse nationale zender

RelFinder zoekt naar alle paden tussen geselecteerde entiteiten waarna men deze kan filteren.

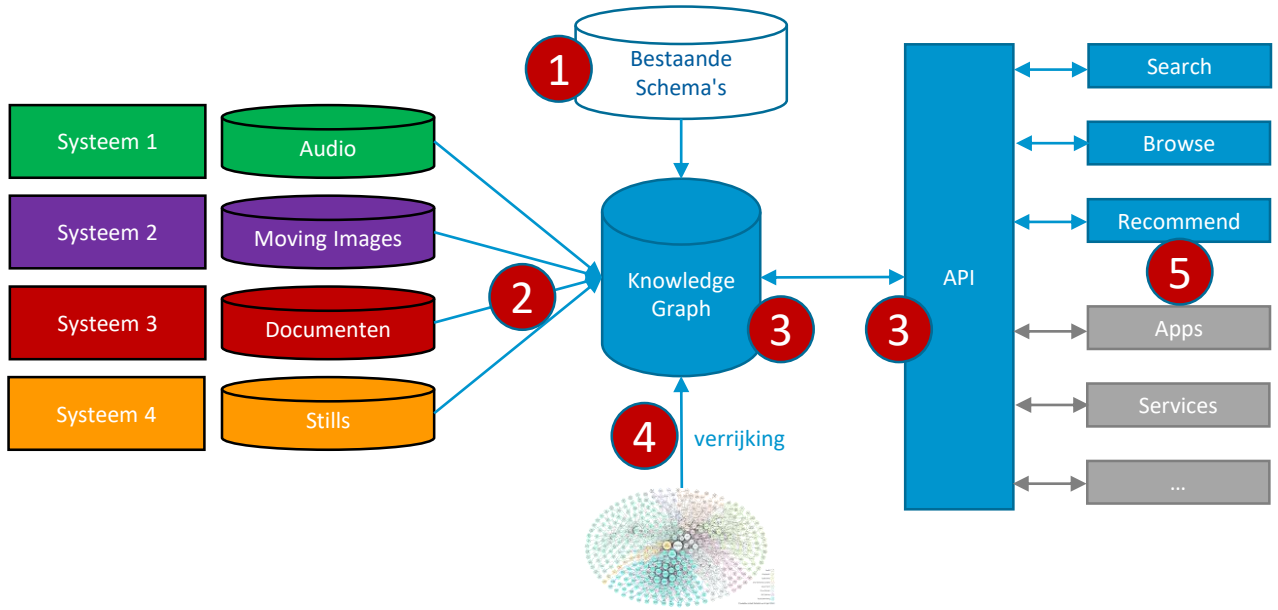
Een voorbeeld van het gebruik van een bestaande applicatie voor knowledge graphs.



RelFinder: http://www.visualdataweb.org/relfinder.php

Smals ICT for society

RTÉ De Ierse nationale zender



1 informaticus en 1 domein expert (tezamen voor 1.5 VTE)
1 jaar voor de proof-of-concept
3 maanden voor 3 systemen, 1 maand voor de laatste
Het corrigeren van de originele data + richtlijnen was een belangrijke taak

Technische Details:

1. Gebruik bestaande schema's
2. Vertaalslag databases naar KG met R2RML; een gestandaardiseerd "taaltje" dat de transformatie beschrijft
3. Apache Jena voor zowel de opslag van de KG en de API
4. De verrijking aan de hand van SILK; een "interlinking framework". De interlinks werden in afzonderlijke "containers" geplaatst
5. Een combinatie van bestaande vrije en eigen tools bovenop de KG

R2RML - https://www.w3.org/TR/r2rml/
R2RML - https://github.com/chrdebru/r2rml
Apache Jena - https://jena.apache.org/
SILK - http://silkframework.org/

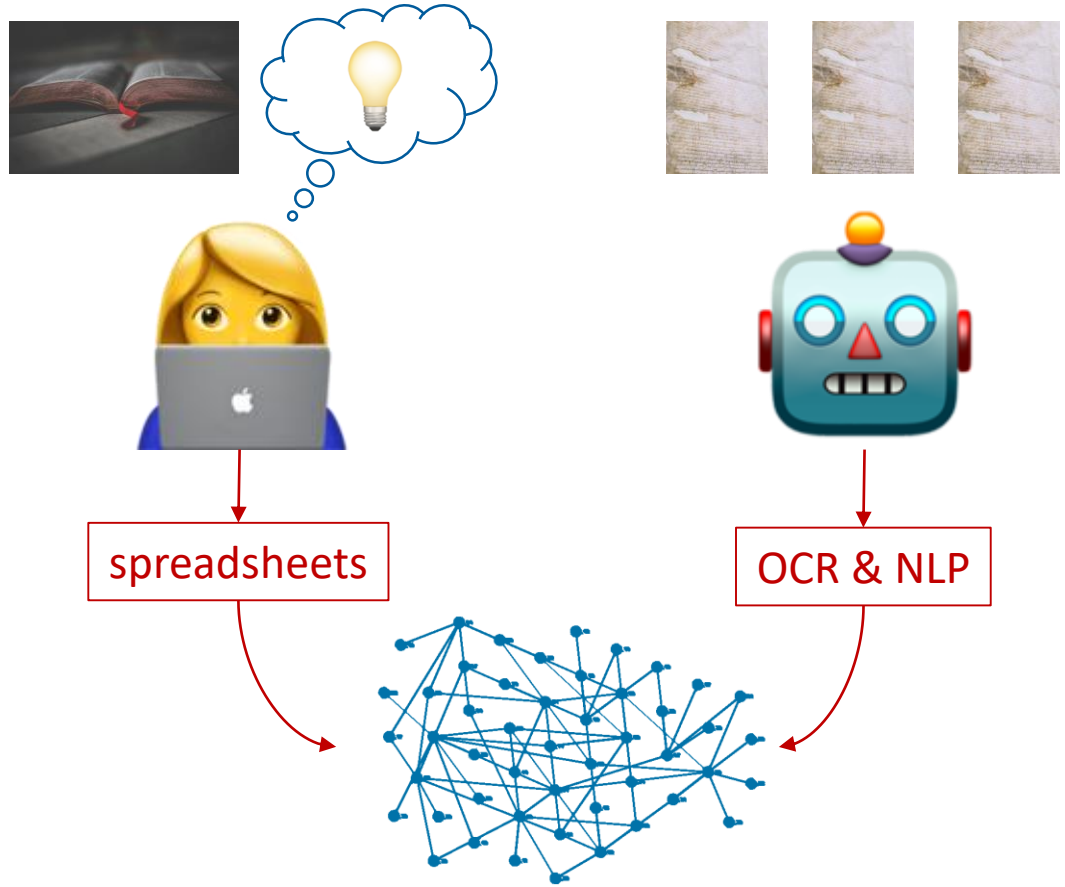
Smals ICT for society

Beyond 2022

Historische personen, mandaten, organisaties, en plaatsen voor een virtueel archief.

Structureren, interpreteren en integreren van informatie uit boeken, brieven, … zowel manueel als automatisch.

Het doel was verder te gaan dan een "gewoon" documentenarchief door het maken van relaties (met context) over documenten heen.



BEYOND 2022
Ireland's Virtual Record Treasury

http://www.beyond2022.ie/

Smals ICT for society

Beyond 2022

Historische personen, mandaten, organisaties, en plaatsen voor een virtueel archief.

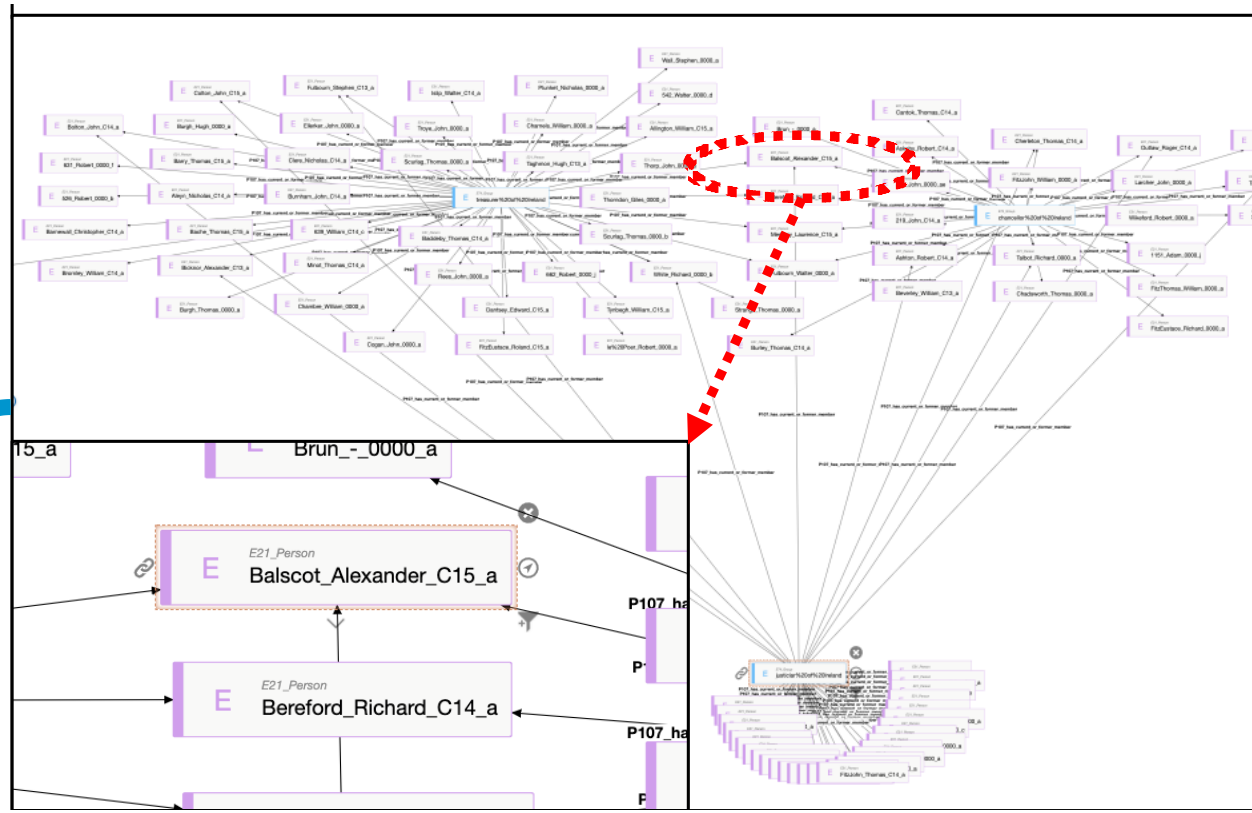
Structureren, interpreteren en integreren van informatie uit boeken, brieven, … zowel manueel als automatisch.

Het doel was verder te gaan dan een "gewoon" documentenarchief door het maken van relaties (met context) over documenten heen.

Bestaande (open) tooling kan met de KG overweg!

Het vinden van de persoon die alle posities hield verliep sneller dan het doornemen van de corpus. Enkele minuten ipv enkele uren!

Doorzoeken van de KG met Ontodia.



https://metaphacts.com/ontodia

Smals ICT for society

Beyond 2022

Technische details:

https://protege.stanford.edu/
https://www.w3.org/TR/shacl/
https://blazegraph.com/

- Bestaande "basis schema" + uitbreidingen met Protégé
- Vertaalslag spreadsheet met R2RML (of Python waar redelijk)
  - Spreadsheets herbruikbaar voor verschillende scenario's
- Kwaliteitscontrole aan de hand van SHACL – een taaltje voor logische- en vormcontroles op KG's
- Data opslag en API met blazegraph
- Een combinatie van bestaande vrije en eigen tools bovenop de KG

Effort voor de KG (zonder NLP pipeline):

- Hoofdzakelijk 1 ingenieur (20-40%) en 1 domeinexpert (80%)
- Regelmatig sessies met andere experten ("focus" en "trial" groepjes)
- 1 jaar voor de **methodologie en proces** en het verwerken van 2 bronnen
- Sustainability primeerde sinds het begin, alsook een agile en incrementele manier van werken

Het bouwen en het onderhouden van een KG

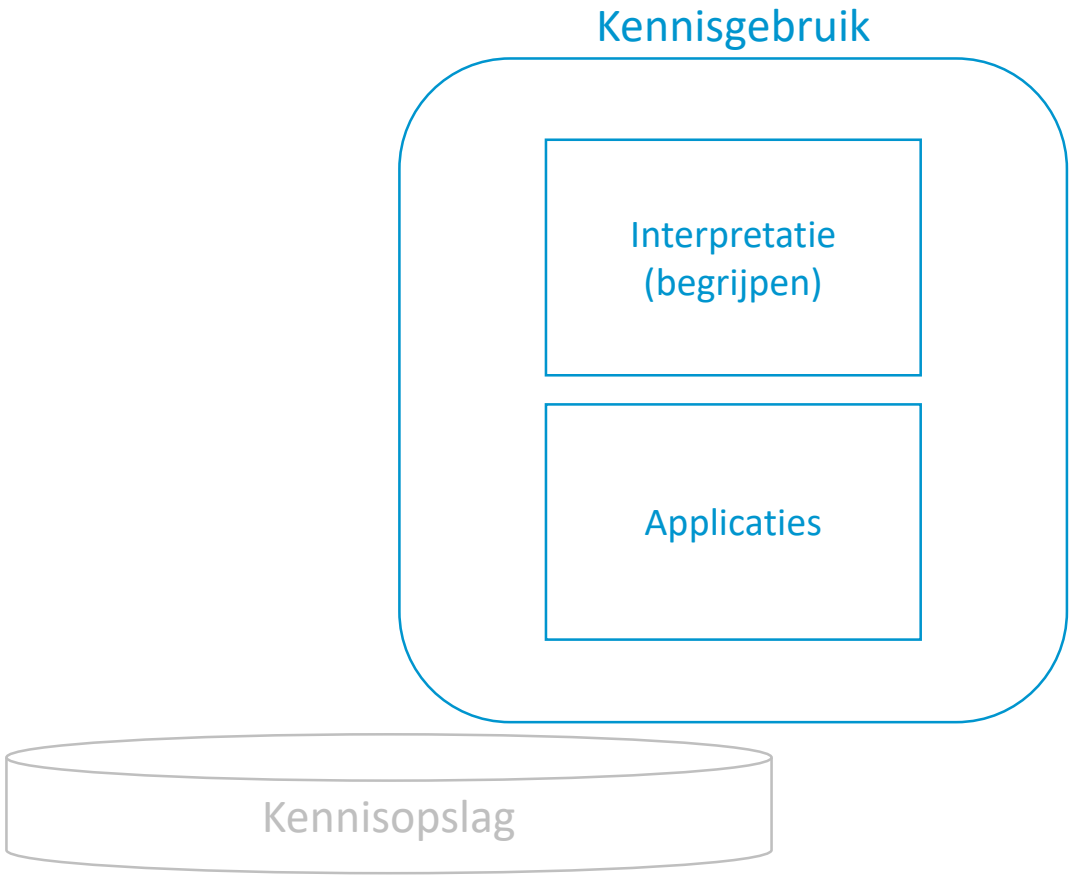
Wat komt er allemaal bij te kijken?

Het opslaan van een KG

Figuur gebaseerd op
Denaux et al., 2017

Smals
ICT for society

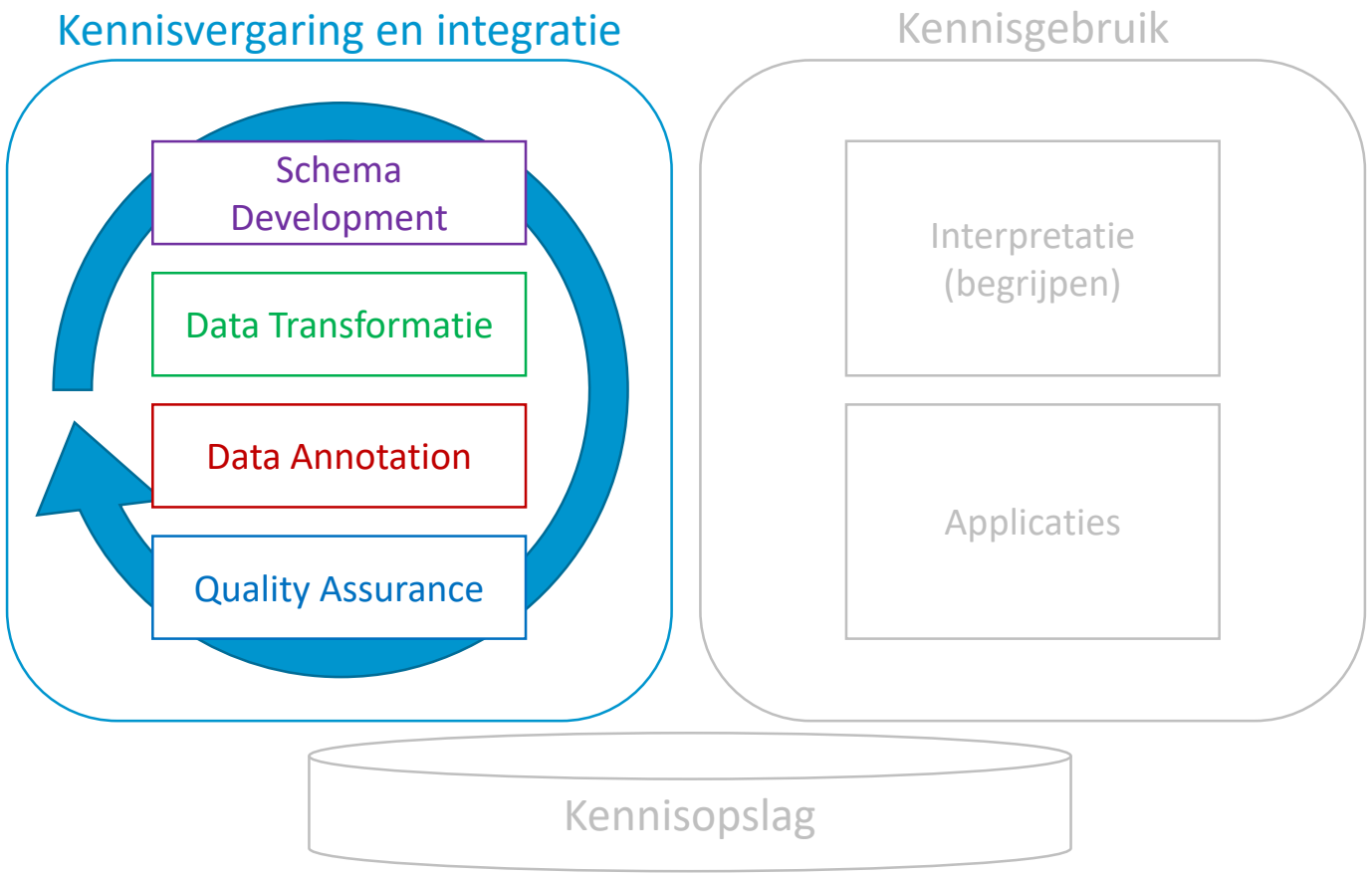
Het gebruiken van een KG



Figuur gebaseerd op
Denaux et al., 2017

Smals
ICT for society

Maken, verrijken, verfijnen, en beheren



Figuur gebaseerd op Denaux et al., 2017

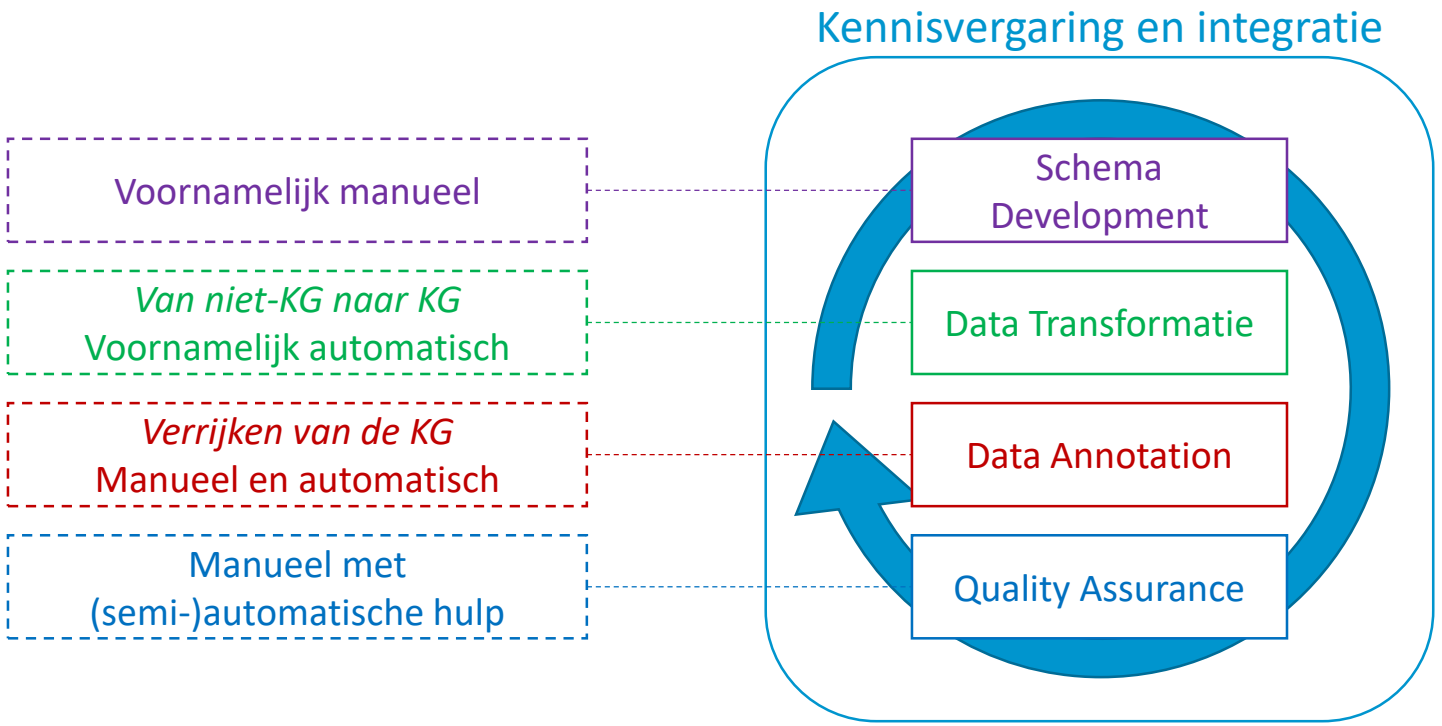
Smals ICT for society

Maken, verrijken, verfijnen, en beheren

Het bouwen en onderhouden van een KG is niet triviaal

Er bestaan voor elke activiteit tools en technieken

Welke tools en technieken adequaat zijn, hangt af van de use case



Klein beginnen en, gedreven door de noden en vereisten, de KG laten evolueren en groeien

Figuur gebaseerd op
Denaux et al., 2017

Smals
ICT for society

Uitdagingen

Automatisch

- Volledigheid
- Consistentie

Manueel

- Beschikbare expertise
- Werkbelasting
- Samenwerking en samenwerkingsproces

Gemeenschappelijk

- Vereisten formuleren
- Evaluatie van de KG
- Expressiviteit vs. efficiëntie
- Mate van hergebruik
- Privacy
- …

Onnauwkeurigheden zijn doorgaans tolereerbaar voor opzoekingen, maar niet voor het nemen van beslissingen.

Uitdagingen

Het project, de use case, en vereisten bepalen in grote mate

De omvang en kost voor het maken van de KG

- Welke zijn de kwaliteitscriteria?
- De "gezaghebbendheid" van de KG?
- Is het eenmalige grote constructie?
- Mate van automatisatie?
- …

De omvang en kost voor het onderhouden van de KG

- **Wie neemt de taak op?**
- **Wie gaat de mensen trainen, opleiden, …?**
- Hoe vaak moeten we annoteren?
- Moeten modellen regelmatig hertraind worden?
- Eenvoudige correcties vs. governance modellen?
- …

"Bezint eer ge begint" – Men moet *minstens* over die vragen denken alvorens een project te starten.

Tools en technieken

Vrije componenten: (niet exhaustief)

- Schema Development
  - Stanford (Web) Protégé
- Data Transformation
  - R2RML engines
- Data Annotation
  - Link discovery: SILK
- Quality Assurance
  - SHACL: TopBraid, Apache Jena
- Data opslag
  - Apache Jena, Blazegraph

Commerciële oplossingen: (niet exhaustief)

- PoolParty
- TopBraid EDG
- Stardog
- Neo4j
- …

Bouwen, onderhouden, gebruiken en opslaan van een KG?
Gebruiken en koppelen van componenten vs. commerciële oplossingen die verschillende aspecten ondersteunen

Dus

Mijn tip is om klein te beginnen

Een duidelijke scope of probleem

Is of zal een KG nuttig zijn?

De voorwaarden van een KG kunnen helpen

Problemen, noden, etc. prioriteren en op een incrementele en "agile" manier werken

Vermijden van premature optimalisatie

Voorbeelden

RTÉ	→ Eerst "Search" en "Browse", en dan "Recommend"
B2022	→ Visie was duidelijk, maar stappen niet. Iteratief concept-per-concept de KG uitwerken

Samenvatting

Een toelichting van het concept knowledge graph: graaf + voorwaarden

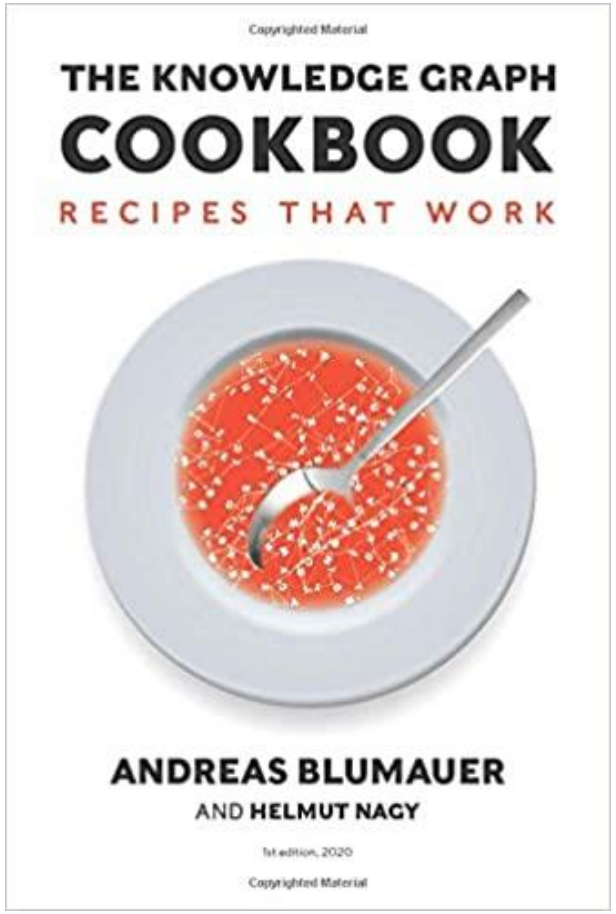
KG's voorzien een flexibele manier om informatie op een betekenisvolle manier te representeren, te integreren, te delen, en te gebruiken

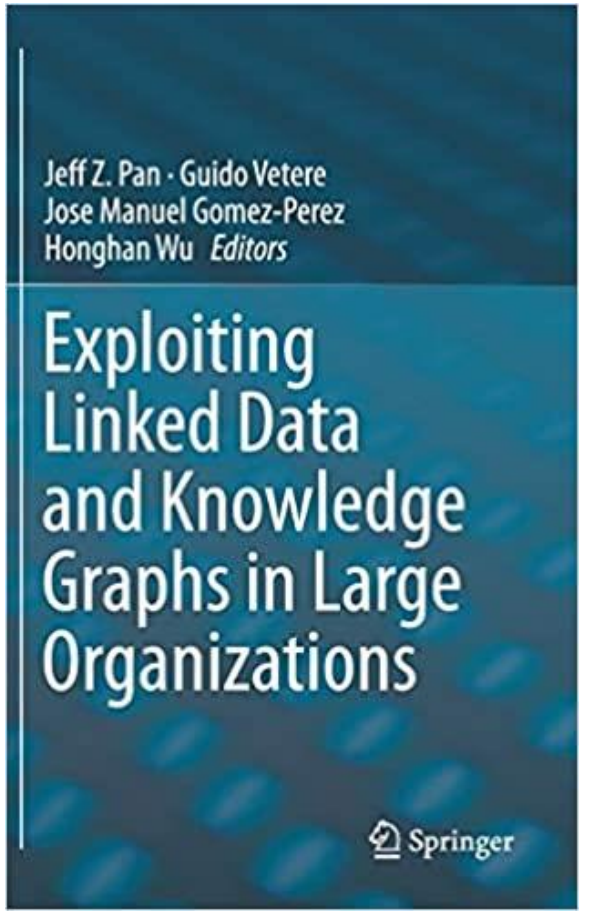
Het concept is simpel, maar de constructie en het onderhouden van een KG is niet triviaal

Wat er bij het onderhouden van een KG komt kijken wordt vaak onderschat

Kans op slagen groeit met realistische verwachtingen en een duidelijke scope

Bibliografie





Smals
ICT for society

Opnieuw een ballonnetje

Semantic Search & Information Retrieval

Recommender Systems

Kennisretentie

Chatbots en Virtual Assistants

Business en Financial Analytics

Meertalige diensten

Operational Research (logistics)

Drug Discovery

…

Smals Research kan u helpen te kijken of een KG een goede oplossing kan zijn voor uw case.

Welke soort problemen?

Een oplossing omvat:

- Bewaren en toegankelijk maken van kennis
- Afleiden van kennis en inzichten

Uit informatie van

- Verschillende domeinen
- Verschillende bronnen
- Niet expliciet opgeslagen
- …

Problemen en applicaties hoeven niet groot te zijn!

Bronnen

Smals Research

Katy Fokou: Les graphes de connaissance, incontournable pour l'intelligence artificielle

Katy Fokou: Les graphes de connaissance : quelques applications

Christophe Debruyne: SHACL: Logische- en vormcontroles met kennisgraaftechnologieën

Literatuur

R. Denaux, Y. Ren, B. Villazón-Terrazas, P. Alexopoulos, A. Faraotti, H. Wu: Knowledge Architecture for Organisations. Exploiting Linked Data and Knowledge Graphs in Large Organisations 2017: 57-84

A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A.-C. Ngonga Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, J. F. Sequeda, S. Staab, A. Zimmermann: Knowledge Graphs. CoRR abs/2003.02320 (2020)

J. Z. Pan. Constructing and Understanding Knowledge Graphs. Tutorial at the 4th Joint International Semantic Technology Conference (JIST 2014). November 9, 2014. Chiang Mai, Thailand.

J. Z. Pan, G. Vetere, J. M. Gómez-Pérez, H. Wu: Exploiting Linked Data and Knowledge Graphs in Large Organisations. Springer 2017, ISBN 978-3-319-45652-2

Bedankt!

Christophe Debruyne

christophe.debruyne@smals.be

Website en nieuwsbrief:

https://www.smalsresearch.be/

Ideeën voor een klein project of een beperkte proof-of-concept?

research@smals.be

Smals
ICT for society