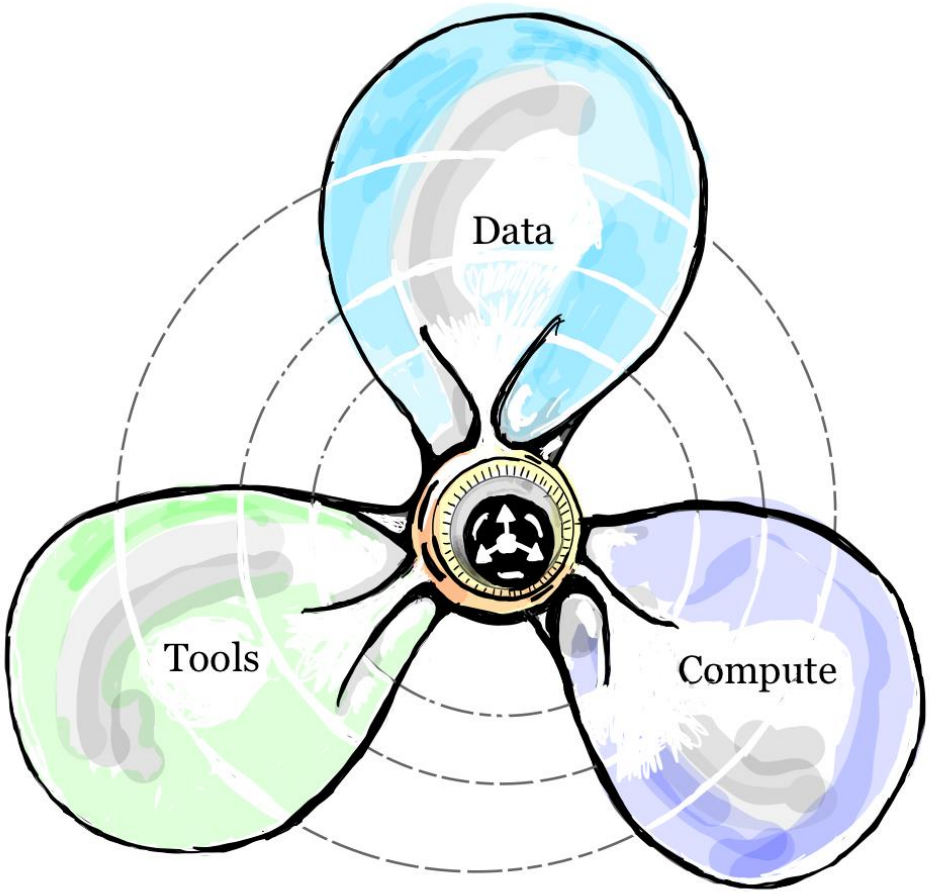
lessons learned in pioneering semantic technology for drug discovery

Barend Mons
Stockholm, PSWC

What we need: The Internet of FAIR data and Services



https://vimeo.com/162062013

Pre-history


BioMed Central
Login
Search B...
BMC Bioinformatics
HOME
ABOUT
ARTICLES
SUBMISSION GUIDELINES
COMMENTARY
OPEN ACCESS
Which gene did you mean?
Barend Mons
BMC Bioinformatics 2005 6:142 DOI: 10.1186/1471-2105-6-142 © Mons; licensee BioMed Central Ltd. 2005
Received: 26 May 2005 Accepted: 07 June 2005 Published: 07 June 2005
Abstract
Computational Biology needs computer-readable information records. Increasingly, meta-analysed and pre-digested information is being used in the follow up of high throughput experiments and other investigations that yield massive data sets. Semantic enrichment of plain text is crucial for computer aided analysis. In
Made it to be my most hated statement
Download PDF
Export citations
Table of Contents
Abstract
Text mining? .....Why bury it first and then mine it again?
Too much to read
Semantics: a crucial addition to text
Databases and pathway tools are not enough
However, text is a nightmare for

History


Nano-Publication in the e-science era
Barend Mons and Jan Velterop
The Anatomy of a Nano-publication
Paul Groth
Andrew Gibson
Johannes Velterop
COMMENTARY
The value of data
nanopub: a beginner's guide to datapublishing
The Open PHACTS Nanopublication Guidelines
The Open PHACTS RDF/Nanopublication Working Group
V1.8 14-2-2012
Open PHACTS
Open Pharmacological Space
LUMC
EURETOS
Knowledge
Discovery
Nanopub.org
Nov
2009
2010
Mar
Apr
2011
Sep
Feb
Apr
2012
Concept
Web
Alliance
The Melton Foundation

The FAIR principles became the unifying force

something to refer to

The FAIR Guiding Principles for scientific data management and stewardship

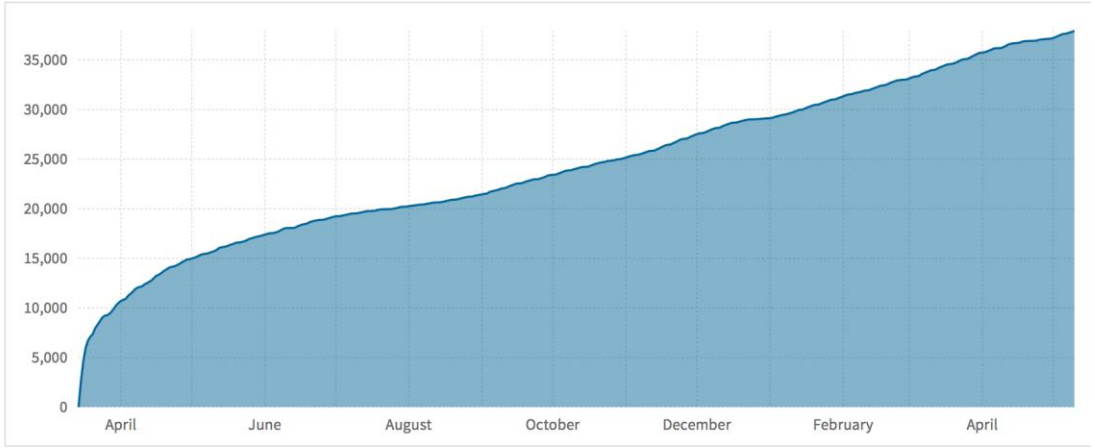
MD Wilkinson, M Dumontier, IJJ Aalbersberg, G Appleton, M Axton, ...
Scientific data 3

114

2016

Page views (37,928)

View as: Cumulative | Line | Table



Online attention

Altmetric score (what's this?)

- Tweeted by **756**
- Blogged by **19**
- On **13** Facebook pages
- Mentioned in **5** Google+ posts
- Picked up by **1** news outlets

Show more

This Altmetric score means that the article is:

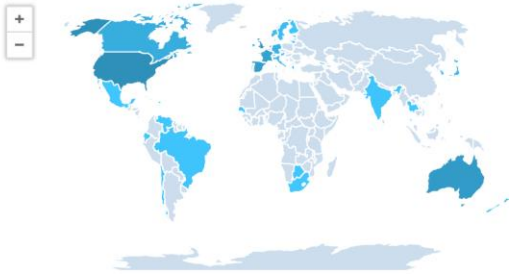
- in the 99th percentile (ranked 615th) of the 278,235 tracked arti similar age in all journals
- in the 95th percentile (ranked 1st) of the 23 tracked articles of a in *Scientific Data*

Mentions in news, blogs & Google+

News articles (1) | Scientific blogs (19) | Google+ posts (5)

Guidelines for best practices in the publication of scientific data
Phys.org

Twitter demographics




G20 HANGZHOU SUMMIT
HANGZHOU, CHINA 4-5 SEPTE
'We support appropriate efforts to promote open science and facilitate appropriate access to publicly funded research results on findable, accessible, interoperable and reusable (FAIR)'

Open Science

Open Innovation

FAIR data from a pharma perspective

from: Herman van Vlijmen (Janssen and chair Open FACTS Foundation)

Open PHACTS

- A lot of data in certain domains
- A lot of it is “unFAIR”, “triple-F (Far From FAIR)”
- Extensive use of CROs makes data standards essential
- Internal development of FAIR data is in early stages
- Strong interest in **FAIR data concept** because of opportunities for answering complex scientific questions and multidomain use cases
  - Open PHACTS (IMI)
  - Open Targets application (EBI)
  - FAIRification of IMI/EFPPIA data (new IMI call topic)
- Key driver of effort should be USE CASES

EMBL-EBI

The European Bioinformatics Institute
Industry Programme

efpia
European Federation of Pharmaceutical Industries and Associations

Affiliate Members

Open PHACTS (IMI project 2011-2016)

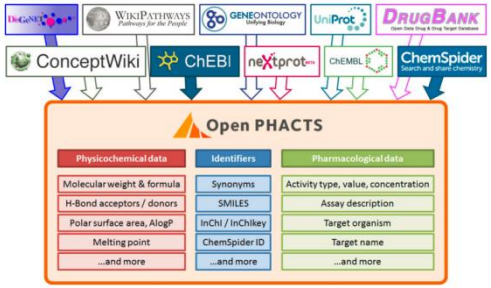
Open PHACTS

Objective: Integrate multiple research biomedical data resources into a single open, sustainable and free access point

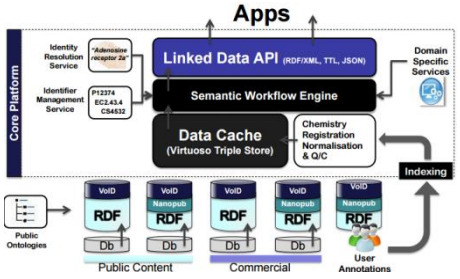
Pharma partners: Pfizer, GSK, Lundbeck, Lilly, Janssen, AZ, Novartis, Esteve

Pharma needs:

- **Data Integration**: public data is heavily underused and needs processing to make it usable with internal data
- **Tools** to answer more complex questions that go across scientific domains



Integrated databases



Available to the world

Scientific competency questions as the basis for semantically enriched open pharmacological space development

Kamal Azzaoui[1], Edgar Jacoby[14], Stefan Senger[2], Emiliano Cuadrado Rodríguez[3], Mabel Loza[3], Barbara Zdrazil[4], Marta Pinto[4], Antony J. Williams[5], Victor de la Torre[6], Jordi Mestres[7], Manuel Pastor[7], Olivier Taboureau[8], Matthias Rarey[9], Christine Chichester[10], Steve Pettifer[11], Niklas Blomberg[12,a], Lee Harland[13], Bryn Williams-Jones[13] and Gerhard F. Ecker[4]

Scientific questions

Answerable with query tools and workflows

Priority use cases collected in Open PHACTS project

TABLE 1

The top 20 research questions

Question number	Question
Cluster I	
Q1	Give me all oxidoreductase inhibitors active <100 nM in human and mouse
Q2	Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?
Q3	Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives
Q4	For a given interaction profile – give me similar compounds
Q5	The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X
Q6	A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature
Q7	Give me all active compounds on a given target with the relevant assay data
Q8	Identify all known protein–protein interaction inhibitors
Q9	For a given compound, give me the interaction profile with targets
Q10	For a given compound, summarize all 'similar compounds' and their activities
Q11	Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)
Cluster II	
Q12	For my given compound, which targets have been patented in the context of Alzheimer's disease?
Q13	Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?
Q14	Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What

Scientific competency questions as the basis for semantically enriched open pharmacological space development

Drug Discovery Today • Volume 18, Numbers 17/18 • September 2013

Kamal Azzaoui[1], Edgar Jacoby[14], Stefan Senger[2], Emiliano Cuadrado Rodríguez[3], Mabel Loza[3], Barbara Zdrazil[4], Marta Pinto[4], Antony J. Williams[5], Victor de la Torre[6], Jordi Mestres[7], Manuel Pastor[7], Olivier Taboureau[8], Matthias Rarey[9], Christine Chichester[10], Steve Pettifer[11], Niklas Blomberg[12,a], Lee Harland[13], Bryn Williams-Jones[13] and Gerhard F. Ecker[4]

Open PHACTS

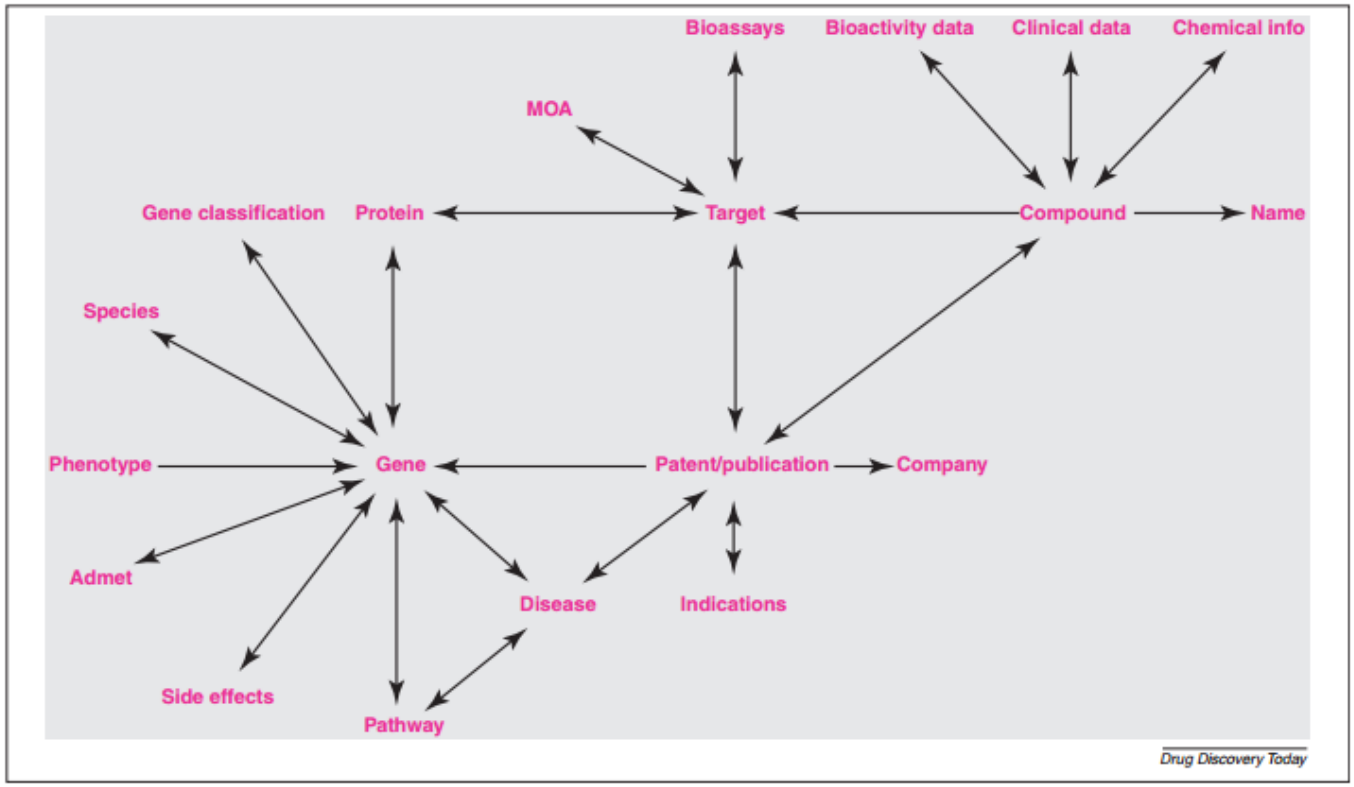
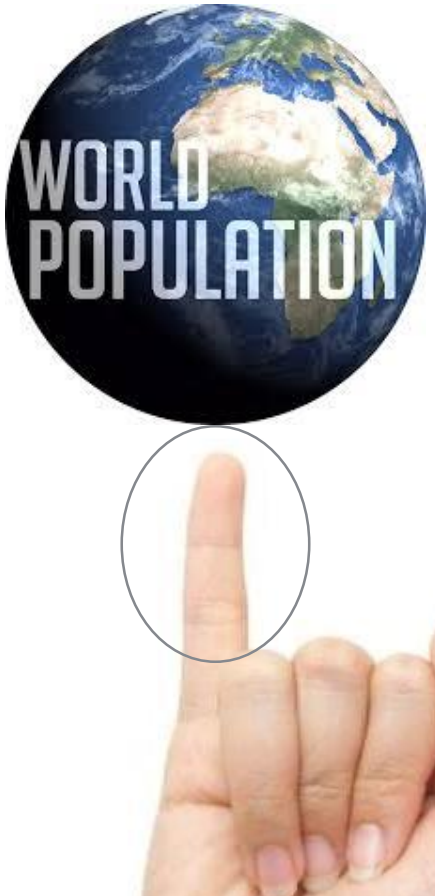


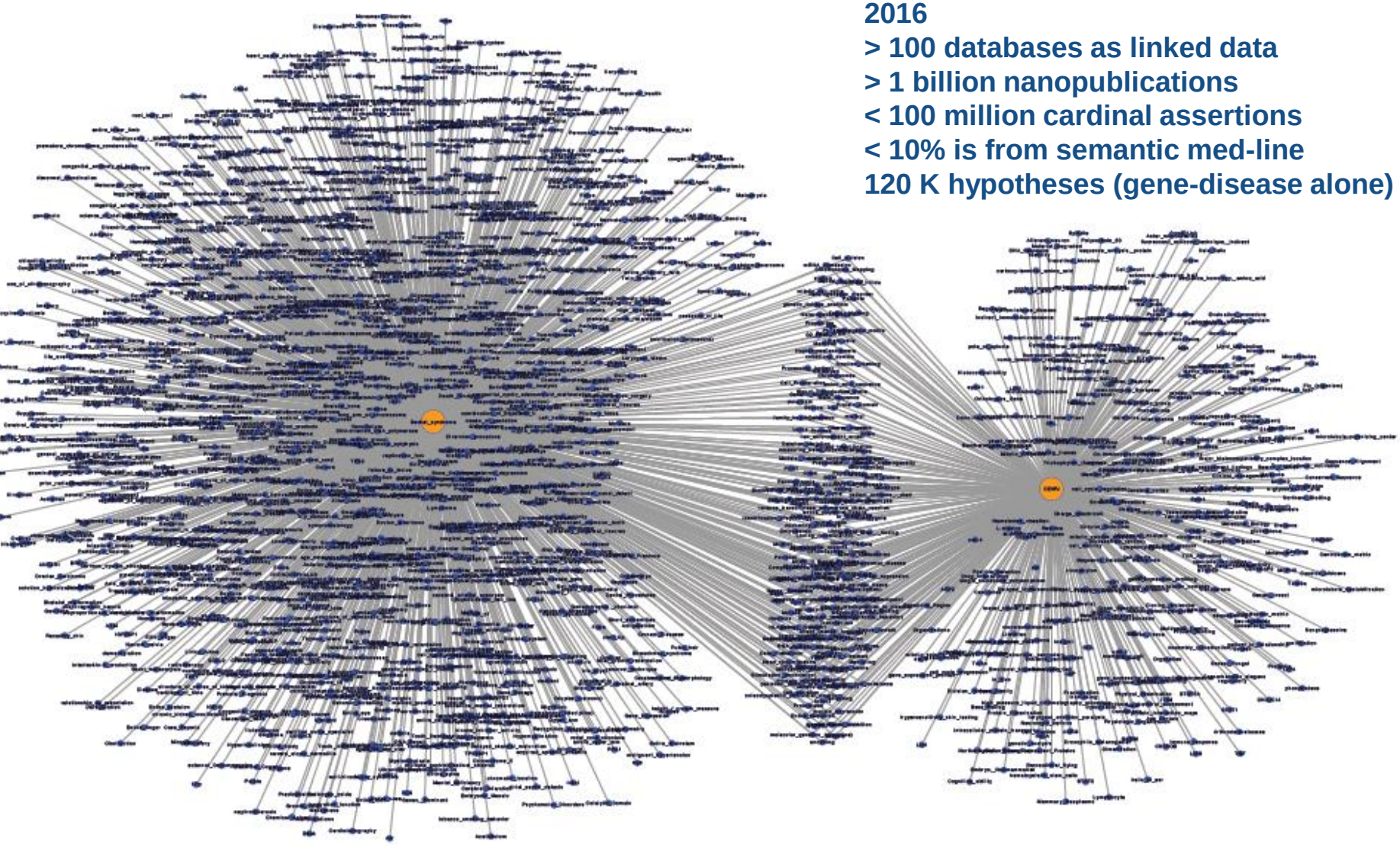
FIGURE 2

Network of data associations needed to answer the top-ranked scientific competency questions. The network reflects a cartoon that summarizes the data associations that are needed to target the top 20 research questions.


ROC
RDF
AUC
HTTP://xzuw
1010 101010
111010 1111
11111010111
ATCG
TCAGAAAT
GCAATTCGTAAT
Phenotype drug
mitochondrion
TT

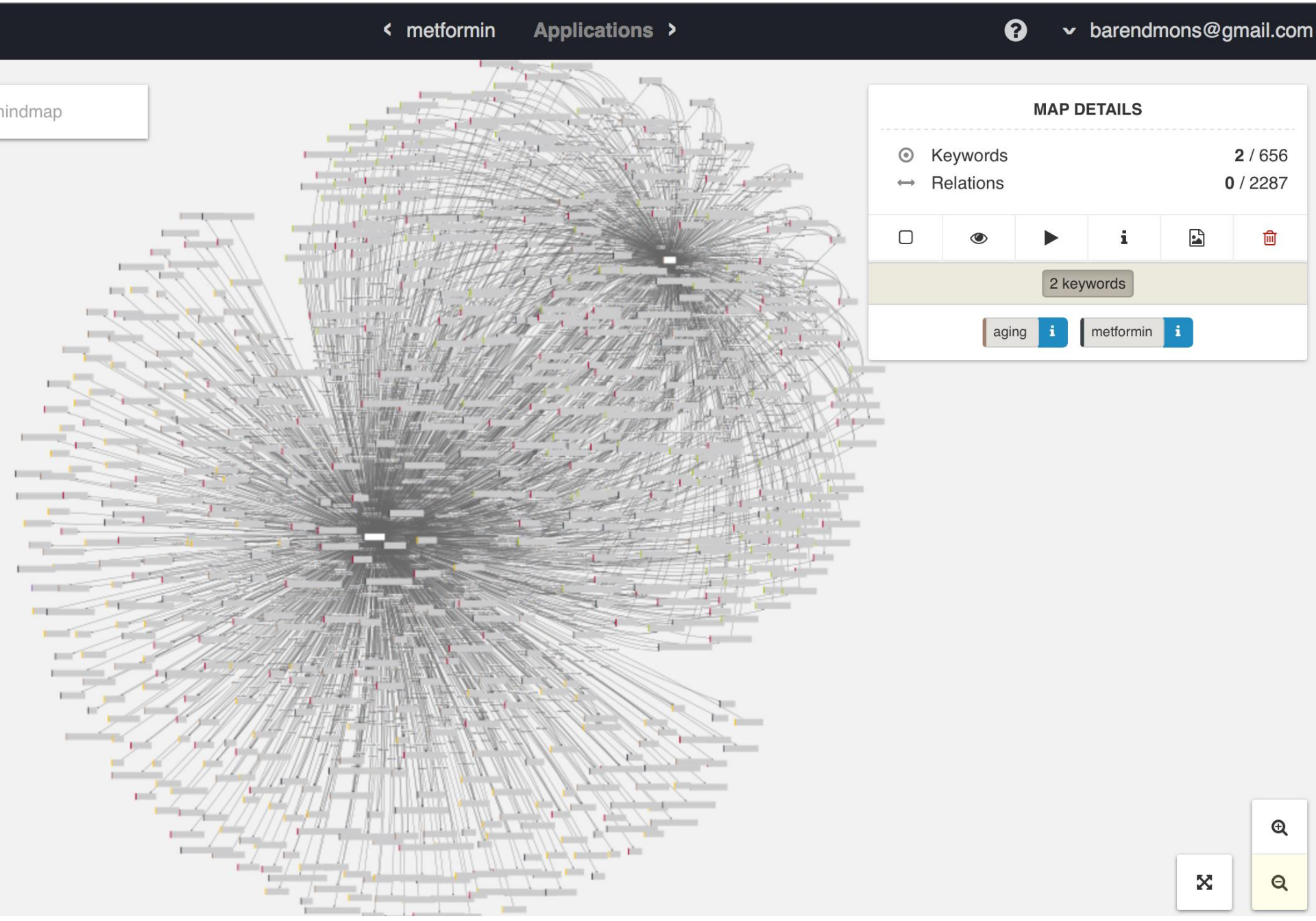

WORLD
POPULATION

We publish about less than a million LSConcepts !

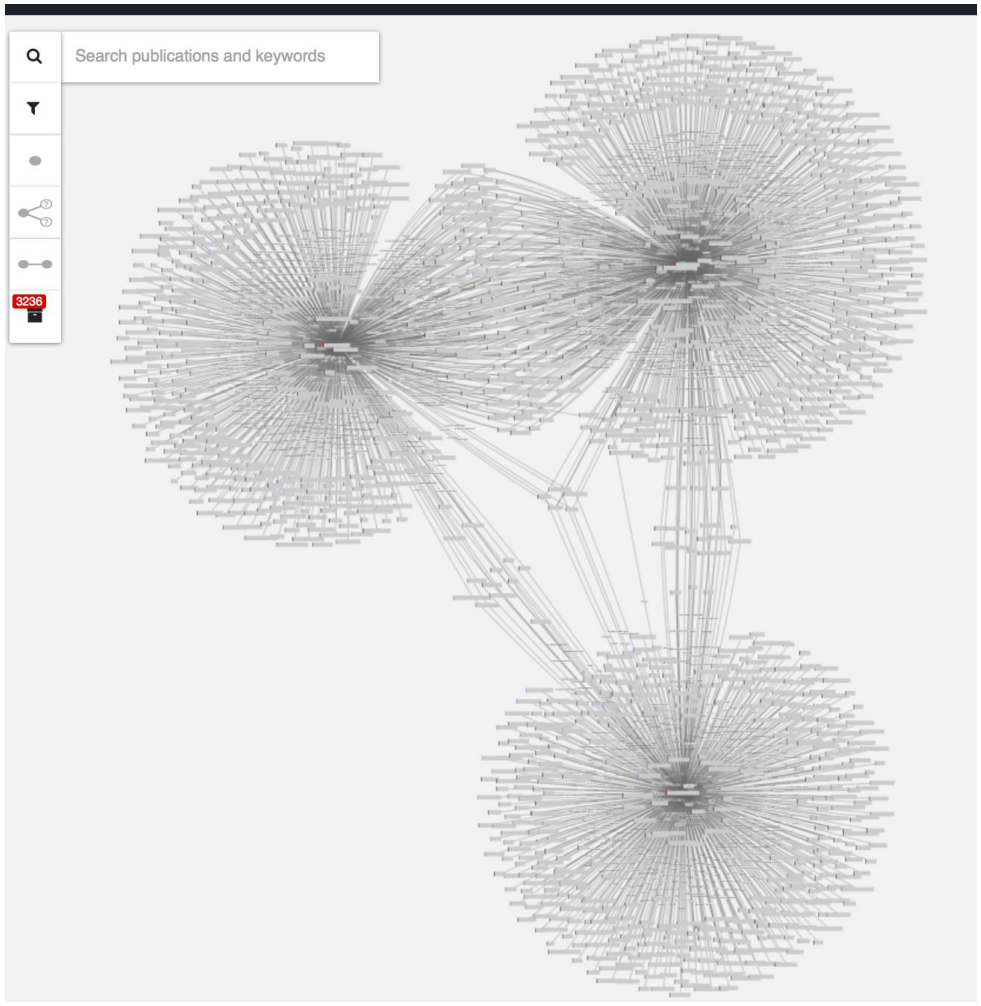


2016
> 100 databases as linked data
> 1 billion nanopublications
< 100 million cardinal assertions
< 10% is from semantic med-line
120 K hypotheses (gene-disease alone)

Made during Nir's talk on Metformin



A real knowlet image (Euretos Knowledge Platform)



share: in this case 5 objects are shared between all three knowlets (in this case: metabolic syndrome


NARRATIVE

Summary Article

Genes predicted for [disease] based on known genes.

Abstract

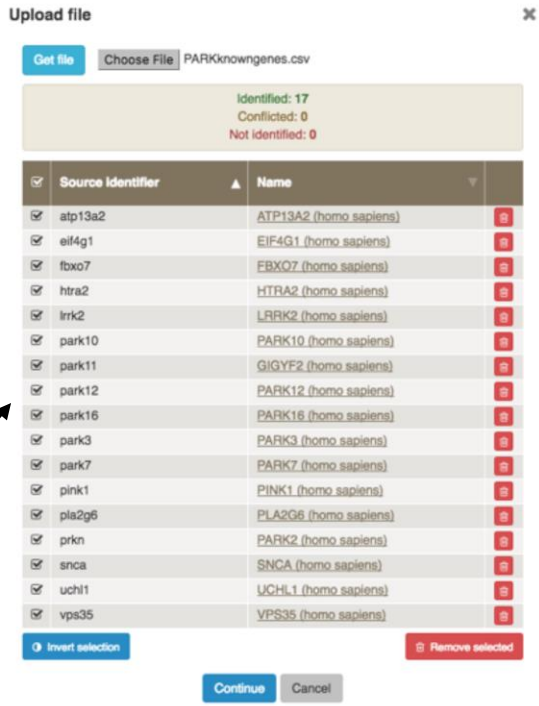
Many genetic diseases, even those that are thought to be caused by mutations in a single gene do show a wide variation in disease phenotypes, severity of symptoms and for instance response to drugs and other interventions. These variation may be partly due to variations in other genetic factors. Not only variants in other associated genes, but also regulation factors and differences in metabolism and general physiology, which in turn may have a strong genetic basis. Also age factors and environmental factors may play a role.
Here, we have used [x] genes that are previously reported to be involved in the [disease] and closely related phenotypic disorders such as [subdisorder 1, sub disorder 2, closely related phenotype] to predict candidate genes that may have a direct of indirect influence in the phenotypic manifestation of [disease] patients.
[x] candidate genes were found with a prediction score ranging between [x] and [y] [reference to mothership paper to be written about methods used in workflows]
For each of those genes we have recovered all relevant direct and indirect associations between the gene, other genes in the cluster and relevant physiological and molecular processes.

Method
The detailed methods used for recovery and analysis sof direct and indirect associations found in [x] pre-mined and annotated, interoperable data resources are described in [ref. 1]. Briefly, all databases were etc.

Results:

Summary (table comparable to tabel in seed article)

main disease	parkinson disease	more information	viewer
sub-phenotypes included	x,y,z	HPO	
number of genes used as reference set	17	nature genetics (DOI)	
predicted genes with direct connections	x	show	show
predicted genes	y	show	show
most frequent connecting semantic categoreis			
other genes and proteins	x	show	show
gene inhibiting molecules	x	show	show
pathways	x	show	show
cellular processes	x	show	show
chemicals, metabolites and drugs	x	show	show
tissues	x	show	show
cell types	x	show	show
organelles	x,y,z	show	show

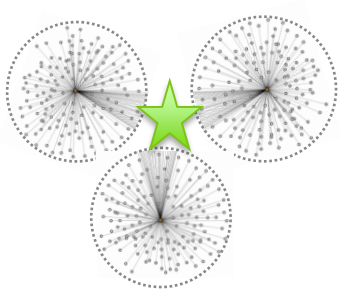



H-factor
Journal impact factor
Tenure track
Nature, science....
Text, Tables, Figures
F.A.I.R

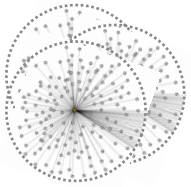

Discount on:
Spurious correlations
Ridiculograms
Not in stock:
Reproducible results
Actionable knowledge
ATGGCPG ATT AA
S
P
O
Biomarker
indicates
condition
Lorum
random
ipsum
HUNT
AI2
PARK
€ $
1
2
H
3
REPOSiSTORIES
Genuine professorwate!
Loss guaranteed!


Supplementary
DATA
PLEASE MINE
PAY HERE
ARTICLES
open science
SPARQ
TAVERNA
F.A.I.R
RAILWAYS

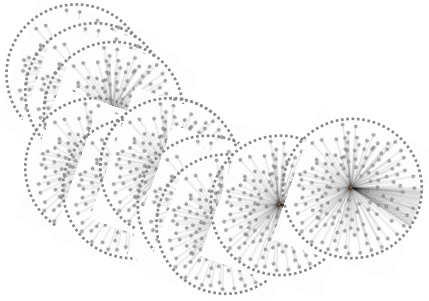
near-sameness and semantic drift of concepts represented by Knowlets



1. The implicitome > new hypothesis generated by the computer

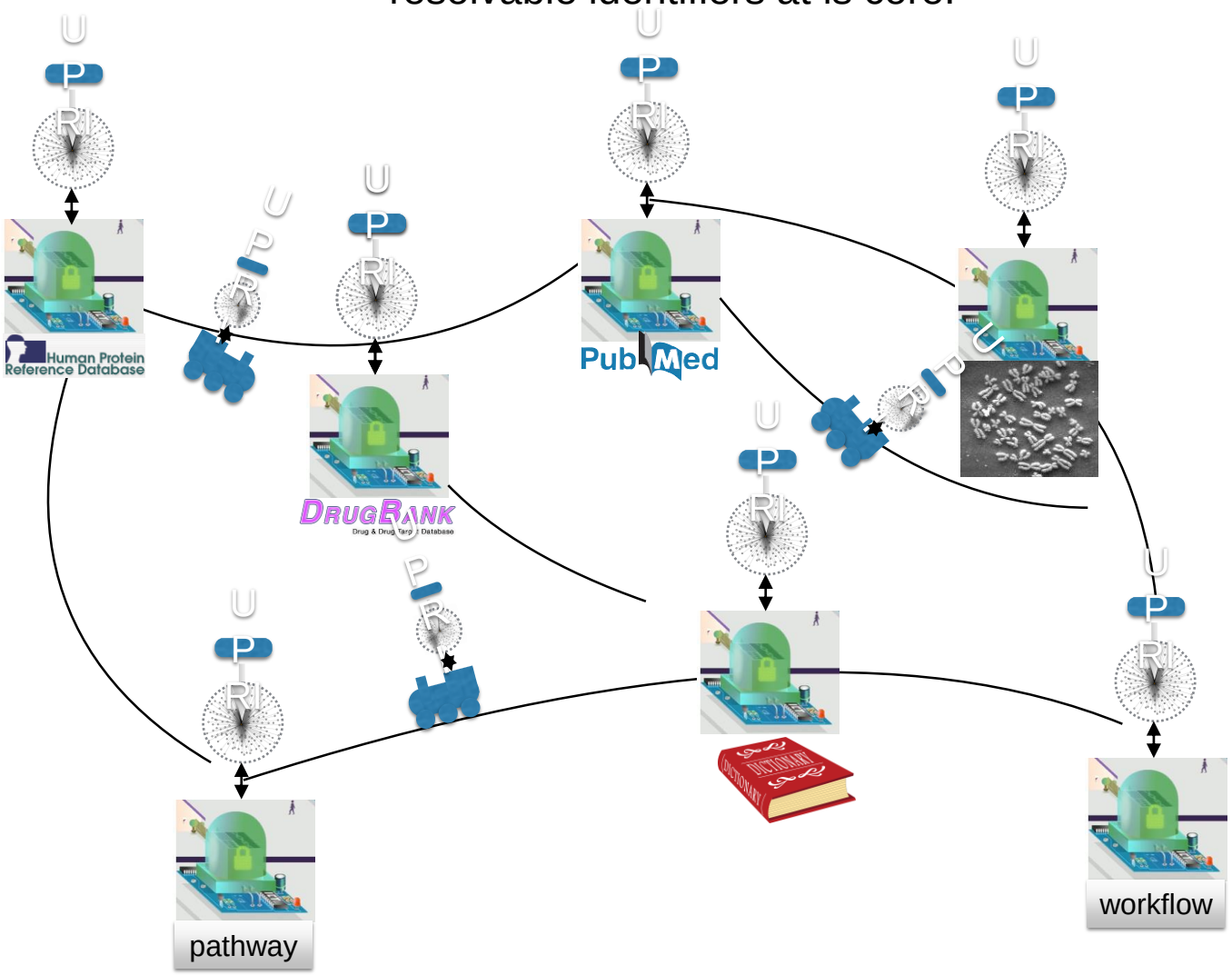


2. dealing with near-sameness (human versus mouse gene etc.)



2. dealing with semantic and conceptual drift)

The Internet of Data and Services (PHT model)
controlled at two levels: UPRI's and a web of associative Knowlets representing abstract concepts, digital objects (including distributed VMs) and physical objects, devices & things in a universal way with scalable, unique, resolvable identifiers at is core.



PSWC2022, Stockholm October 18th , in combination with FORCE 2020
Pre-symposium event:
The Nobel Prize in NON-Literature
The Swedish Academy (Svenska Akademien), Börssalen, Källargränd 4, Stockholm
More information: http://svenskaakademien.se/
The first Nobel prize ever awarded to a Social Machine !! (join us)
Reception Co-sponsor: The Innovative medicines Initiative