Surveillance Graphs

Jonny L. Saunders

March 24, 2023

Contents

1. toc {:toc}

Subpages:

- Outline
- TODO
- Scraps

{% include acknowledgements.html %}
{% include foreword.html %}

1 Introduction

The world is Big Data, and The Cloud is its landlord. Much like the universal commodification of prior [1] assemblages of capitalism, The Cloud's informational capitalism recasts every element of our reality as Data, and it is our responsibility to ferret it out of its primitive unknown, mine it, harvest it, dump it by the tanker-truckful into great Data Lakes and wring the Actionable Insights from its neck. The world has always been Big with Data, but The Cloud has finally given us a fruit of knowledge to munch on that reveals its true nature. We are but Data Subjects, and if we can harness the wily spray of our Organic Content with biosensor wearables, filtering our every action, affection, and affilitation through a mob of algorithmically optimized

platforms, then The Cloud might teach us enough about ourselves to be able to lead a fulfilled and healthy life. Academics and Governments in particular are filthy with Data, and have set their eyes on a new generation of data infrastructures that promise to dissolve the Silos that prevent the Big Data from teaching us the nature of the universe, how to solve poverty, reverse climate change, and finally collapse into one great cat-pile of peace and love for our neighbor. The Cloud is dreaming of linking the Big Data into one great Knowledge Graph of Everything — it tells us this is important for the fate of humanity.

The Knowledge Graph of Everything is a mirage, though. Its vision of extracting all the world's data is the same colonial vision of infinite prosperity that has driven us to the brink of extinction: So The Cloud Said, let us make Platforms in our image, and let them have dominion over the interactions with the apps, and the insights of the analytics, and over all the data of the earth. The pursuit of universalizing ontologies that will finally give every quantum of Data its one True and Correct form is the same fascistic vision that has driven eliminationist campaigns to stamp out degeneracy: Ontologien über alles. Aside from its properly pathological nature, the Knowledge Graph of Everything *is impossible* and *won't work.* Instead, The Cloud will lead governments and academics along by the nose just long enough to build critical mass for an interlocking set of platforms that slice off snapshots of the Everything to ratchet us ever futher into the captivity of surveillance and subscription.

What is the alternative? The Cloud presents its future as inevitable, but by seeing past its logic we might imagine properly *human* infrastructures that fill the needs for connection and understanding that it exploits. This piece develops the notion of **vulgar linked data** as an alternative to the Cloud Orthodoxy. Predicated on relationality, heterogeneity, privacy, and vernacular expression, vulgar linked data infrastructures attempt to empower *people* to *socially organize* information in a truly decentralized sociotechnological commons, rather than empowering *systems* to *rent* knowledge organization for *profit.*

2 Knowledge Graphs: A Backbone in the Surveillance Economy

Through their cloud of corporate jargon, knowledge graphs are relatively straightforward to define [2, 3, 4, 5] (though see [6]): **directed, labeled graphs** consisting of *nodes* corresponding to entities like a person, dataset, location, etc. and *edges* that describe their relationship[1]. Knowledge graphs typically make use of some controlled **ontology** that provide a specific set of terms and how they are to be used, and "types" that give a given entity an expected set of *properties* denoted by edges with a particular set of labels from the ontology. For example, the schema.org Person type would be applied to a node, and then have a set of labeled edges like `gender` or `email` that link to other nodes that contain the values of the properties.

Why does such a seemingly ordinary data structure deserve particular attention in an always-more-fraught landscape of digital technology? The story of knowledge graphs is the story of the enclosure of the wild and open web into a series of surveillance-backed platforms. They provide an underexplored lens onto the present and future of digital infrastructure as planned by information conglomerates — and serve as a liberatory kernel that hints at how we might chart a different course.

2.0.1 Semantic Web: Priesthoods

The term "Knowledge Graph" evolved out of the Semantic Web project [3] . It is difficult to reconstruct how radical the notion of a collection of documents organized by arbitrary links between them was at dawn of the internet. At the time, the infrastructures of linking documents looked more like ISBNs, carefully regulated by expert, centralized authorities[2]. Being able to *just make anything that could be linked to* and *link to anything you wanted* was *terrifying* and *new* (eg. [8, 9]).

The initial design of the web imagined it as a self-organizing process, where people would maintain their own websites and organize a collection of links to other websites. It became clear relatively quickly that the anarchy of a socially self-organizing internet wasn't going to work as planned, where without a formal system of organization "people were frightened of getting lost in it. You could follow links forever." [10]

[1] Equivalently, one could emphasize that they are graphs composed of **triplet** links that describe some subject, predicate, and object.

[2] For another example re: the political nature of the DOI system in the face of the arbitrary linking of the internet, see section 3.1.2 "Integration, not Invention" in [7]

In its earliest formulations, the Semantic Web was an attempt to supplement the same arbitrary power to express human-readable information with computer-readable information. It imagined a linked and overlapping set of schemas ranging from locally expressive vocabularies used among small groups of friends through globally shared, logically consistent ontologies. The semantic web was intended to evolve fluidly, like language, with cultures of meaning meshing and separating at multiple scales [11, 12, 13] :

> Locally defined languages are easy to create, needing local consensus about meaning: only a limited number of people have to share a mental pattern of relationships which define the meaning. However, global languages are so much more effective at communication, reaching the parts that local languages cannot. [...]
>
> So the idea is that in any one message, some of the terms will be from a global ontology, some from subdomains. The amount of data which can be reused by another agent will depend on how many communities they have in common, how many ontologies they share.
>
> In other words, one global ontology is not a solution to the problem, and a local subdomain is not a solution either. But if each agent has uses a mix of a few ontologies of different scale, that is forms a global solution to the problem. [11]

> The Semantic Web, in naming every concept simply by a URI, lets anyone express new concepts that they invent with minimal effort. Its unifying logical language will enable these concepts to be progressively linked into a universal Web. [12]

This freeform goal expression for expression's sake was always in tension with another part of the vision - serving as a backbone for AI "agents" that could compute emergent function from the semantic web. Succinctly: "Human language thrives when using the same term to mean somewhat different things, but automation does not." [12] This tension persists through the broader history of the web.

2.0.2 Linked Data: Platforms

Much of the work of the semantic web project in the early 2000s focused on the "global" side of this tension at the expense of the "local" - creating ontologies and related technologies intended to serve as a foundation for expressing basic things in a common vocabulary [3] . This work had many successes, but began a schism between the priesthood of people concerned with making systems that were *correct* and those that were more concerned with making things that *worked* - or supported "local" expression (eg [14]). Aaron Swartz captured this frustration in his unfinished book:

> Instead of the "let's just build something that works" attitude that made the Web (and the Internet) such a roaring success, they brought the formalizing mindset of mathematicians and the institutional structures of academics and defense contractors. They formed committees to form working groups to write drafts of ontologies that carefully listed (in 100-page Word documents) all possible things in the universe and the various properties they could have, and they spent hours in Talmudic debates over whether a washing machine was a kitchen appliance or a household cleaning device. [15]

Lindsay Poirier describes this difference in "thought styles" as a rift between the "neats" focused on universalizing *a priori* ontologies and the "scruffies" focused on everyday use and letting the structure appear afterwards [16] . The latter characterizes the "second age" of the Semantic Web after 2006 - the reorganization around **Linked Data** [17, 3] . The era of Linked Data de-emphasized the idealistic and ideological goals of the early Semantic Web, driven more by an empirical approach of trying to realize these systems on the wilds of the web, creating some of the first public "Linked Open Data" systems like DBPedia and Freebase.

This turn coincides with the emerging platformatization and enclosure of the web as "Web 2.0." Throughout the early 2000s, the work of the Semantic Web project was largely invisible to the ordinary web user, and its vision of a self-organizing web was easily outcompeted by the now-ubiquitous use of search engines to index

the web. Where in the early 2000s web architects were imagining the future of web continuing to take place on free and open *protocols,* the Linked Data/Web 2.0 era corraled us into a pattern of *platforms* which quickly ratcheted their way to dominance in a positive feedback loop of user experience design, network effects and profit. On platforms, rather than a system that "belongs" to everyone, you are granted access to some specific set of operations through an interface so that you can be part of a social process of producing and curating information for the platform holder.

2.0.3 Knowledge Graphs: Panoptica

In 2010 Google acquired Metaweb and its publicly-edited Semantic Web database Freebase, and in 2012 repackaged it and the ideas of Linked Data as what it called a **Knowledge Graph** — the third era of the Semantic Web [18, 19] . Freebase only made up part of it, and the full extent of Google's Knowledge Graph are unknown, but its most visible impact are the factboxes that present structured information about the subjects of searches - like biographical information in a search for a person - or the different widgets for contextual interaction - like being able to make a restaurant reservation from the serach page [20] . Knowledge Graphs still share the same underlying structure — triplet graphs with ontologies — even if they occupy a broader space of implementations and technologies. What differs is the context and intended use: the "worldview" of the knowledge graph.

Beyond the obvious product-level features it supports, Google's acquisition of Freebase and the structure of its Knowledge Graph represent at least two deeper shifts in the trajectory of the Semantic Web and the broader internet: the privatization of technologies with initially liberatory aspirations, and an early template of the sprawling, surveillance-driven information conglomerate we know and love today.

Like the radical nature of linking on the web, it's difficult to remember that the web as surveillance apparatus thinly veiled as the five or so remaining websites was not inevitable. The pre-dotcom bust internet of the 90's and early 2000's was far from the commercialized wasteland we know today. Ed Horowitz, CEO of Viacom explained in 1996: "The Internet has yet to fulfill its promise of commercial success. Why? Because there is no business model" [21] . Google's AdWords being a defining moment in the development of surveillance capitalism is a story already told [22] : taking advantage of the need for search generated by the disorganization of the web, AdWords turned personal search data into a profit vector by selling targeted space in the results.

The significance of the relationship between search, the semantic web, and what became knowledge graphs is less widely appreciated. The semantic web was initially an alternative to monolithic search engine platforms - or, more generally, to platforms in general [23] . It imagined the use of triplet links and shared ontologies at a protocol level as a way of organizing the information on the web into a richly explorable space: rather than needing to rely on a search bar, one could traverse a structured graph of information [17, 24] to find what one needed without mediation by a third party.

Instead, the form of of the semantic web that emerged as "Knowledge Graphs" flipped the vision of a free and evolving internet on its head. The mutation from "Linked Open Data" [17] to "Knowledge Graphs" is a shift in meaning from a public and densely linked web of information from many sources to a proprietary information store used to power derivative platforms and services. The shift isn't quite so simple as a "closure" of a formerly open resource — we'll return to the complex role of openness in a moment. It is closer to an *en*closure, a *domestication* of the dream of the Semantic Web. A dream of a mutating, pluralistic space of communication, where we were able to own and change and create the information that structures our digital lives was reduced to a ring of platforms that give us precisely as much agency as is needed to keep us content in our captivity. Links that had all the expressive power of utterances, questions, hints, slander, and lies were reduced to mere facts. We were recast from our role as *people* creating a digital world to *consumers* of subscriptions and services. The artifacts that we create for and with and between each other as the substance of our lives online were yoked to the acquisitive gaze of the knowledge graph as *content* to be mined. We vulgar commoners, we data subjects, are not allowed to touch the graph — even if it is built from our disembodied bits.

The same technologies, with minor variation, that were intended to keep the internet free became emblematic of and coproductive with the surveillance/platform model that has enclosed it. Beyond Google, knowledge graphs are an elemental part of the contempory information economy. Banks, militaries, governments, life

science corporations, journalists, everyone is using knowledge graphs [25, 26] . Their ubiquity is not an accident, one of many possible data systems that could have fit the bill, but reflects and reinforces basic patterns of the information economy and the corporations within it.

What makes knowledge graphs so special? It turns out that semantic web technologies, designed to accomodate the infinitely heterogeneous, multiscale nature of free and unmediated social structuring of information are also quite useful for the indefinitely expanding dragnet of data collection that defines the operation of contemporary capitalism:

> "If one takes a look at the top Fortune 500 companies, it is surprising how many of them are really in the information business. I don't just mean the technology and telecommunication companies like Apple or Google or Verizon or Cisco or the drug companies like Pfizer. One could also think of the big banks as a subset of the vectoralist class rather than as "finance capital." They too are in the information asymmetry business. And as we learned in the 2008 crash, even the car companies are in the information business—they made more money from car loans than cars. The military—industrial sector is also in the information business. The companies that appear to sell actual things, like Nike, are really in the brand business. Walmart and Amazon compete with different models of the information logistics business. Even the oil companies are in part at least in the information-about-the-geology-of-possible-oil-deposits business. Perhaps the vectoralist class is no longer emerging. Maybe it is the new dominant class." - *McKenzie Wark, Capital Is Dead: Is This Something Worse?* [1]

Data companies — most major companies — need to store and maintain massive collections of heterogeneous data across their byzantine hierarchies of executives, managers, and workers. This gigantic haunted ball of data is not just a tool, but the *substance* of the company. A data company persists by exploiting the combinatorics of its data hoard, spinning off new platforms that in turn maintain and expand access to data by creating captive data subjects[3]. As it expands, a conglomerate will acquire many new sources and modalities of data and need to integrate them with its existing data.

Knowledge graphs are particularly well suited for this "data integration" problem. A full technical description is out of scope here, but briefly: traditional relational database systems can be very difficult to modify and refactor, and that difficulty increases the larger and more complex a database is[4]. One has to be design the structure of the anticipated data in advance, and the abstract schematic structure of the data is embedded in how it is stored and accessed. It is particularly difficult to do unanticipated "long range" analyses where very different kinds of data are analyzed together.

In contrast, merging graphs is more straightforward[5] [2, 26, 27, 28, 29, 30, 31, 32] - the data is just triplets, so in an idealized case[6] it is possible to just concatenate them and remove duplicates (eg. for a short example, see [33, 34]). The graph can be operated on locally, with more global coordination provided by ontologies and schemas, which themselves have a graph structure [35] . Discrepancies between graphlike schema can be resolved by, you guessed it, making more graph to describe the links and transformations between them. Long-range operations between data are part of the basic structure of a graph - just traverse nodes and edges until you get to where you need to go - and the semantic structure of the graph provides additional constraints to that traversal. Again, a technical description is out of scope here, graphs are not magic, but they are well-suited to merging, modifying, and analyzing large quantities of heterogeneous data.

[3] Facebook describes the notion of its platform as being just a means of interacting with its underlying data graph in corporate web design speak: "A useful tool for Facebook has been to think of the graph as the model and a Facebook page as the view—a projection of an entity or collection of entities that reside in the graph." [20]

[4] For a practical example, see a recent trio of blog posts from Etsy engineers that describe the process of scaling their database system.

[5]

> That is because knowledge graphs aim to solve the data incongruence problem, which is one of the biggest operational headaches for corporates, says Atkin. "Corporates suffer from technology fragmentation and as a result have a lot of data that doesn't align across the organization. Doing the hard work to fix this data incongruence reality is a pre-requisite for realizing business value," he says. [27]

[6] I am aware graph databases are not magic and this is an extraordinarily simplified example. The principle is the point, not all the subtle ways the implementations of graph databases are hard.

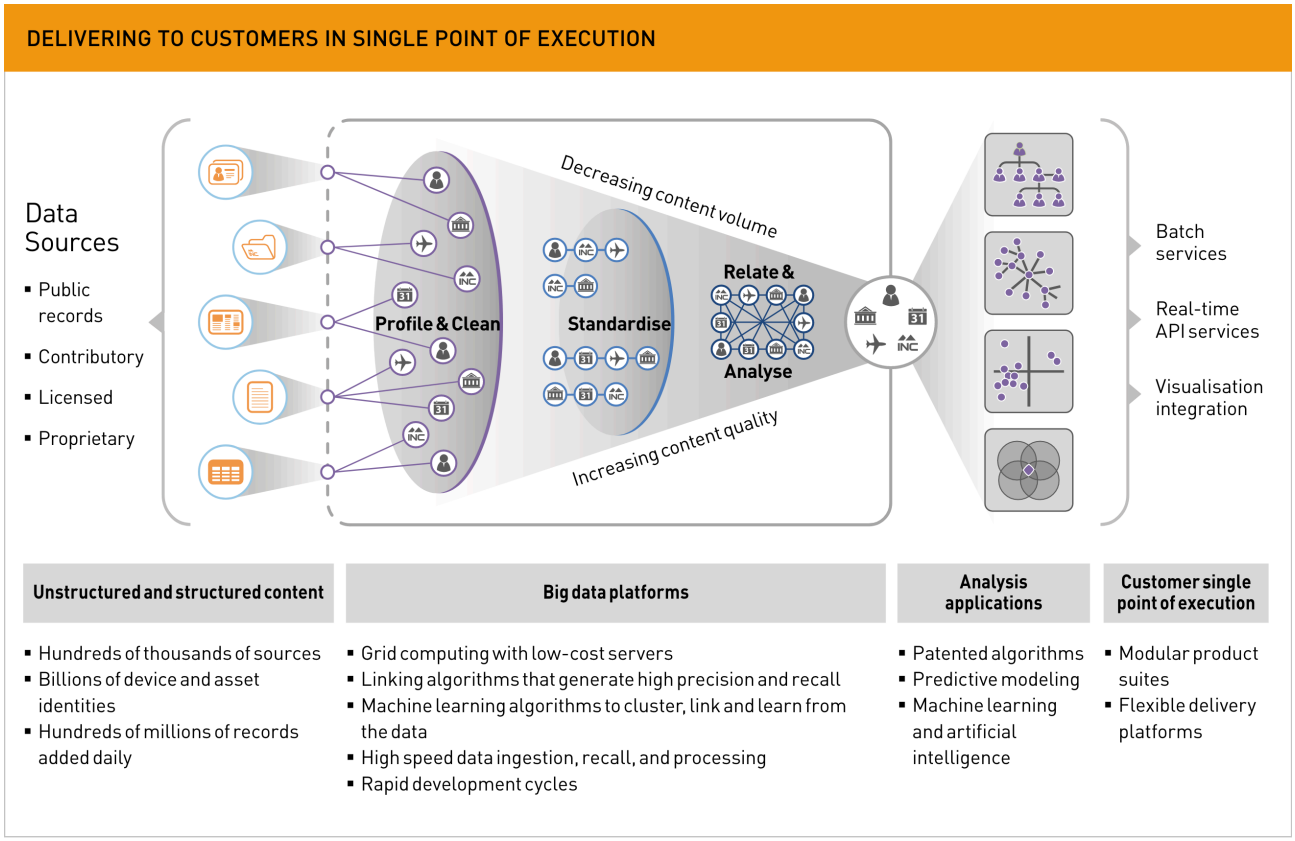
Another way of looking at the capacity for heterogeneity in triplet graphs is by thinking of links as statements:

> One person may define a `vehicle` as having a `number of wheels` and a `weight` and a `length`, but not foresee a `color`. This will not stop another person making the assertion that a given car is `red`, using the color vocabulary from elsewhere. [10]

So if you are a data broker, and you just made a hostile acquisition of another data broker who has additional surveillance information to fill out for the people in your existing dataset, you can just stitch those new properties on like a fifth arm on your nightmarish data frankenstein.

What does this look like in practice? While in a bygone era Elsevier was merely a rentier holding publicly funded research hostage for profit, its parent company RELX is paradigmatic of the transformation of a more traditional information rentier into a sprawling, multimodal surveillance conglomerate (see [36]). RELX proudly describes itself as a gigantic haunted graph of data:

> Technology at RELX involves creating actionable insights from big data – large volumes of data in different formats being ingested at high speeds. We take this high-quality data from thousands of sources in varying formats – both structured and unstructured. We then extract the data points from the content, link the data points and enrich them to make it analysable. Finally, we apply advanced statistics and algorithms, such as machine learning and natural language processing, to provide professional customers with the actionable insights they need to do their jobs.
>
> We are continually building new products and data and technology platforms, re-using approaches and technologies across the company to create platforms that are reliable, scalable and secure. **Even though we serve different segments with different content sets, the nature of the problems solved and the way we apply technology has commonalities across the company.** [37]



In its 2022 Annual Report, RELX describes its business model as ingesting large quantities of data, linking them together, and deriving platforms from them. [37]

While to any individual market segment or class of customers RELX and its subsidiaries might look like a portfolio of separate platforms and applications, one can only make sense of the company by thinking of each of them as a view on an interconnected graph of data[7]. Each additional source of data, either by acquiring new companies or by expanding their existing control of informational access points has the potential to create some combinatorically new set of opportunities for new platforms.

For example, RELX is able to gather surveillance data on researcher attention data through the tracking in its ScienceDirect and Mendeley platforms. It also collects a large amount of chemical data through its control of scientific publishing that it rents access to on its Reaxys platform, which is supplemented by its LexisNexis PatentSight database of patents. So far so normal.

What about the other sides of the multisided market? RELX is able to combine these and other data sources into new product. For pharmaceutical R&D companies, their bespoke Drug Design Optimization services advertise being able to use chemical, disease, and literature-based data to generate a priority list of potential therapeutic targets and drugs, as well as provide "competitive intelligence" about which targets are currently being studied, presumably identified from their ownership of the scientific literature coupled with surveillance data. Since clinicians don't trust pharmaceutical advertisements [39] , Elsevier uses its position as a perceived neutral third party to repackage advertisements as informational systems [40] , "journal-branded webinars," as well as a number of other avenues via its "360 degree advertising solutions" catalogue. So, by combining several data sources and platforms, Elsevier is able to offer pharmaceutical companies recommendations for candidate drugs above and beyond what would be possible with chemical information alone and then advertise their drugs directly to doctors.

Derivative platforms beget derivative platforms. Its integration into clinical systems by way of reference material is growing to include electronic health record (EHR) systems, and they are "developing clinical decision support applications [...] leveraging [their] proprietary health graph" [37] . Similarly, their integration into Apple's watchOS to track medications indicates their interest in directly tracking personal medical data.

That's all within biomedical sciences, but RELX's risk division also provides "comprehensive data, analytics, and decision tools for [...] life insurance carriers" [37] , so while we will never have the kind of external visibility into its infrastructure to say for certain, it's not difficult to imagine combining its diverse biomedical knowledge graph with personal medical information in order to sell risk-assessment services to health and life insurance companies. LexisNexis has personal data enough to serve as an "integral part" of the United States Immigrations and Customs Enforcement's (ICE) arrest and deportation program [41, 42] , including dragnet location data [43] , driving behavior data from internet-connected cars [44] , and payment and credit data as just a small sample from its large catalogue [45] of data aggregated and linked into comprehensive profiles [46] . The contemporary knowledge graph-powered surveillance conglomerate gains its versatility precisely from its ability to span many unrelated domains and deploy new platforms as opportunities present themselves. As new data sources are acquired, the combinatorics of possible surveillance products correspondingly explode.

This pattern is true across the information industry [28] . A handful of representatives from Microsoft, Google, Facebook, eBay, and IBM describe some elements of each of their knowledge graphs in a 2019 paper [20] . Each has different scopes, applications, and interaction with the other data and processing infrastructure at the company, but all emphasize the ability for their knowledge graphs to accomodate change, heterogeneity, conflicting data, inference, and facilitate work by distributed teams due to their self-documenting and modular nature. Neo4j, developers of an eponymous graph database library, describes in one case study among its hundreds of customers how the U.S. Army uses its "connected data" to track its equipment and estimate the cost of some new exploratory imperialism [47] . An analysis of Palantir's hundreds of patents for knowledge graph technology (eg. [48, 49, 50, 51]) describes its ambitions for its knowledge graph:

> There is evidence [...] that Palantir has infrastructural aspirations to become a general classification system for data integration [...] that can be tailored into a universal knowledge graph. [...] Palantir similarly imagines a world where its platform might serve as a "shadow" universal knowledge graph for governments, industries, and organizations. [52]

[7] Though apparently they have had historical difficulty actually getting that integration to work [38] .

Knowledge graphs *as a technology* - like all technologies - are not intrinsically unethical. It is the structure of the capital-K capital-G Knowledge Graph *as a concept* with its *context* that is pathological. They represent the historical trajectory of semantic web ideas and technologies from something that we are intended to use and create directly into privately held data that we can only interact with through platforms. They are coproductive with the corporate and technical structure of surveillance capitalism, facilitating conglomerates that gobble up as many platforms and data sources as possible to stitch them into an expanding, heterogeneous graph of data. We will return to the underlying ideology of the knowledge graph and an alternative in the final section.

In particular, it is their "graph plus compute" structure - where some underlying graph of data is coupled with a set of algorithms and interfaces to view it - that is necessary to understand some of the more counterintuitive motivations of surveillance conglomerates. This structure complicates questions of "openness" versus "proprietariness," one of the deepest loci of criticism of the platformatized web, and provides a different lens on ostensibly "open" or "public" knowledge graph-based infrastructure projects.

3 Public Graphs, Private Profits

3.0.1 Unqualified Openness Considered Harmful

A reader that I am constructing as a straw man for the sake of argument might ask: if the problem is information conglomerates stockpiling a massive quantity of proprietary data and renting use of it, isn't "open data" the answer? And to that reader I would gently shake my head and say a qualified "no."

"Openness," including open source, open standards, and open data, is a subtle tool that can be used both to dissolve and reinforce economic and political power [53] .

Free and open source software, with its noble (and decidedly non-monolithic [54]) goal of creating an ecosystem of free[8] software, is a means by which large information companies can harvest the commons and outsource labor costs [55, 56, 57, 58, 59] . There are countless examples of FOSS developers maintaing software widely used by companies making billions of dollars for little or no compensation - eg. core-js [60] , OpenSSL [61] , leftpad [62] , PLC4X [63] and so on. When an information company releases or supports an open source project it is rarely an act of altruism. The effect is to prevent another company from profiting from a proprietary version of that technology, signal virtue, drive recruitment, and create a centralized point to concentrate donated labor. Microsoft, a famously good actor in software, took this several steps further with GitHub, VSCode, and later Copilot, capturing a large chunk of the software development *process* in order to trick programmers to be the "humans in the loop" refining the neural network to write code and dilute their labor power [64, 65, 66, 67] .

"Peer production" models, a more generic term for public collaboration that includes FOSS, has similar discontents. The similar term "crowdsource[9]" quite literally describes a patronizing means of harvesting free labor via some typically gamified platform. Wikipedia is perhaps the most well-known example of peer production[10], and it too struggles with its position as a resource to be harvested by information conglomerates. In 2015, the increasing prevalence of Google's information boxes caused a substantial decline in wikipedia pageviews [70, 71] as its information was harvested into Google's knowledge graph, and a "will she, won't she" search engine arguably intended to avoid dependence on Google was at the heart of its 2014-2016 leadership crisis [72, 73] . After shuttering Freebase, Google has donated a substantial amount of money to kickstart its successor [74] Wikidata, presumably as a means of crowdsourcing the curation of its knowledge graph [75, 76, 77] .

"Open" standards are yet another fraught domain of openness. For an example within academia, the seemingly-open Digital Object Identifier (DOI) system was concocted as a means for publishers to retain control of indexing research, avoiding the impact of the proposed free repository PubMedCentral and the high overhead of linking documents between publishers[11] (see sec. 3.1.1 in [7]). The nonprofit standards body

[8] "free as in whatever will prevent you from @'ing me about getting some definition of free wrong."

[9] For critical work on crowdsourcing in the context of "open science," see [68] , and in the semantic web see [69] .

[10] I have written about the peculiar structure of Wikipedia among wikis previously, section 3.4.1 - "The Wiki Way" [7]

[11] "The potential benefit of the service that would become CrossRef was immediately apparent. Organizations such as AIP and IOP

NISO's standards for indicating journal article versions [79] and licensing [80] are used by publishers to enforce their intellectual property monopolies and programmatically scour the web to prevent free access to publicly funded information [81] .

Schema.org, a standard intended to be the generic interchange ontology of the web, is another emblem of enclosure of the semantic web. Its introduction at the SemTech 2011 conference was cause for a rare point of agreement[12] between the then-warring maintainers of RDFa and Microformats: "folks, it's wrong for Google to dictate vocabularies, let's not lose sight of that" [82] . Though ostensibly open, its structure and emphases have been roundly criticized, eg. having a eurocentric bias towards commercially valuable information [83] . It encourages website maintainers to embed Schema.org annotations in their pages in exchange for a boost in search rankings — which Google then embeds in its infoboxes, driving down page views. More fundamentally it cements the notion that Linked Data is something that we are only intended to use to make our information more available to some search engine crawler rather than make use of for ourselves: "In general, the design decisions place more of the burden on consumers of the markup" [84] . It encodes the notion that there should be one "neutral" means of representing information for one (or a few) global search engines to understand, rather than for local negotiation over meaning and location. According to the transcribed Q&A after its 2011 announcement, the Google representatives characterized the creation of authoring tools like those created to make creative use of HTML more accessible as a potential "alternative path," but then dismissed the notion of improved tooling as "impossible" [85] .

Clearly, on its own, mere "openness" is no guarantee of virtue, and socio-technological systems must always be evaluated in their broader context: *what is open? why? who benefits?* Open source, open standards, and peer production models do not inherently challenge the rent-seeking behavior of information conglomerates, but can instead facilitate it.

In particular, the maintainers of corporate knowledge graphs want to reduce labor duplication by making use of some public knowledge graph that they can then "add value" to with shades of proprietary and personal data (emphasis mine):

> In a case like IBM clients, who build their own custom knowledge graphs, **the clients are not expected to tell the graph about basic knowledge.** For example, a cancer researcher is not going to teach the knowledge graph that skin is a form of tissue, or that St. Jude is a hospital in Memphis, Tennessee. This is known as **"general knowledge,"** captured in a general knowledge graph. **The next level of information is knowledge that is well known to anybody in the domain**—for example, carcinoma is a form of cancer or NHL more often stands for nonHodgkin lymphoma than National Hockey League in some contexts it may still mean that—say, in the patient record of an NHL player). **The client should need to input only the private and confidential knowledge** or any knowledge that the system does not yet know. [20]

The creation of a collection of more domain-specific ontologies and tooling for ingesting previously unstructured data would allow for a new kind of globally linked knowledge graph ecosystem — making use of a broader range of publicly-available data, as well as facilitating new markets for renting access to interoperable data. Five information conglomerates conclude their joint paper on knowledge graphs accordingly:

> The natural question from our discussion in this article is whether different knowledge graphs can someday share certain core elements, such as descriptions of people, places, and similar entities. [20]

Having such standards be under the stewardship of ostensibly neutral and open third-parties provides cover for powerful actors exerting their influence and helps overcome the initial energy barrier to realizing network

(Institute of Physics) had begun to link to each other's publications, and the impossibility of replicating such one-off arrangements across the industry was obvious. As Tim Ingoldsby later put it, **'All those linking agreements were going to kill us.'"** [78]

[12](Intervening messages in the chat log have been omitted for clarity):

<tantek> Hey Kavi - do you see what you've done here? <tantek> You've gotten a community leader of microformats.org (myself) and chair of W3C RDFa WG to *agree* <edsu> tantek: see, that's progress :) <manu-db> Yes - both RDFa and Microformats communities agree - sky will be falling, next.

effects from their broad use [86, 87] . Peter Mika, the director of Semantic Search at Yahoo Labs, describes this need for third-party intervention in domain-specific standards:

> A natural next step for Knowledge Graphs is to **extend beyond the boundaries of organisations,** connecting data assets of companies along business value chains. This process is still at an early stage, and **there is a need for trade associations or industry-specific standards organisations to step in,** especially when it comes to developing shared entity identifier schemes. [88]

As with search, we should be particularly wary of information infrastructures that are *technically* open[13] but embed design logics that preserve the hegemony of the organizations that have the resources to make use of them. The existing organization of industrial knowledge graphs as chimeric "data + compute" models give a hint at what we might look for in public knowledge graphs: the data is open, but to make use of it we have to rely on some proprietary algorithm or cloud infrastructure.

Unfortunately, that is exactly what at least two US Federal agencies have in mind: the NIH and NSF are both in the thick of engineering cloud-based knowledge graph infrastructures and domain-specific ontologies with all the trappings of technology that fills the stated needs of information conglomerates at the expense of the people it is outwardly intended to serve. We will describe those efforts and their already apparent risks as a way of understanding how these technologies illustrate and reinforce the ideological and practical dominance of the existing corporate informational ecosystem — and to articulate an alternative.

Add note that we are assuming that people are working with the best of intentinos here, adn that it is hard to imagine an alternative system when the existing one is so dominant! These are mostly good people trying to do good things in a system that's rotten.

3.0.2 NIH: The Biomedical Translator

Note:

This section is reproduced from, focuses, and expands on "Linked Data or Surveillance Capitalism?" from [7] .

The NIH's Biomedical Data Translator[14] project was initially described in its 2016 Strategic Plan for Data Science as a means of translating between biomedical data formats:

> Through its Biomedical Data Translator program, the National Center for Advancing Translational Sciences (NCATS) is supporting research to develop ways to connect conventionally separated data types to one another to make them more useful for researchers and the public. [89]

The original funding statement from 2016 is similarly humble, and press releases through 2017 also speak mostly in terms of querying the data – though some ambition begins to creep in. By 2019, the vision for the project had shifted from *translating* between data types into the realm of heterogeneous linkages in some meta-level system for linking and *reasoning* over them.

In their piece "Toward a Universal Biomedical Translator," then in a feasibility assessment phase, the members of the Translator Consortium assert that universal translation between biomedical data is impossible[15] [90]

[13] Go ahead, try and make your own web crawler to compete with Google - all the information is just out there in public on the open web!

[14] Or, just "Translator"

[15]

> First, we assert that a single monolithic data set that directly connects the complete set of clinical characteristics to the complete set of biomolecular features, including "-omics" data, will never exist because the number of characteristics and features is constantly shifting and exponentially growing. [...] We also assert that there is no single language, software or natural, with which to express clinical and biomolecular observations—these observations are necessarily and appropriately linked to the measurement technologies that produce them, as well as the nuances of language. The lack of a universal language for expressing clinical and biomolecular observations presents a risk of isolation or marginalization of data that are relevant for answering a particular inquiry, but are never accessed because of a failure in translation.
>
> Based on these observations, our final assertion is that automating the ability to reason across integrated data sources and providing users who pose inquiries with a dossier of translated answers coupled with full provenance and confidence in the results is critical if we wish to accelerate clinical and translational insights, drive new discoveries, facilitate serendipity,

. The impossibility they saw was not that of conflicting political demands on the structure of organization (as per [53]), but of the sheer numeracy of the data and vocabularies needed to describe them. The risk posed by a lack of a universal "language" was not being able to index all possible data, rather than inaccuracy or inequity[16].

Undaunted by their stated belief in the impossibility of a universalizing ontology, the Consortium created one in their biolink model[17] [93, 92] . Biolink consists of a hierarchy of general[18] classes: eg. a BiologicalEntity like a Gene, or a ChemicalEntity like a Drug. Classes can then linked by any number of properties, or "Slots[19]," like a therapeutic procedure that treats a disease.

Biolink was designed to be a sort of "meta ontology," or a means of mapping different domain-specific biomedical ontologies onto a common vocabulary[20]. This design reflects the structure of the rest of the Translator ecosystem: the interaction with domain-specific ontologies, the kinds of data sources it uses, and the way that end users are expected to interact with the Translator.

As a meta-ontology, biolink is targeted towards "meta data." Rather than accomodating "raw data[21]," biolink is expected to operate at the level of "knowledge," or "generally accepted, universal assertions derived from the accumulation of information" [96] : this procedure treats that disease, this chemical interacts with that one, etc.

The primary way Biolink is used within the Translator is to structure a registry of database APIs, each called a "Knowledge Source." Knowledge Sources use biolink to declare that they are able to provide assertions about a particular set of classes or slots, like drugs that affect genetic expression, which makes them part of the Translator's distributed Knowledge Graph. The Translator project, in this universalizing impulse, recapitulates some of the early beliefs of the Semantic Web updated with some of the techniques of Linked Data. Since acquiring Knowledge is just a matter of creating the right tools rather than a social process, NIH RePORTER shows a series of grants for small councils of experts to create domain-specific ontologies and Knowledge Sources.

This structure strongly constrains who is intended to be able to contribute to the Translator: highly curated biomedical informatics platforms, rather than basic researchers. This, in turn, reflects deeper beliefs about the nature of information within the Translator ecosystem: "knowledge" is not a social, contextual, or dialogical phenomenon, but a "natural resource" that can be mined from information that is "out there." A scientific paper is a neutral carrier of a factual link between entities. The meaning of "translation," in some uses, has shifted from translating *between data formats*, to *"translating information into knowledge"* [90] . This is, of course, the ideology of Big Data: "when heterogeneous networks are connected at a massive scale, new knowledge can be extracted as an emergent property of the network" [97] . The Translator imagines itself as a refinery, converting crude data into knowledge that can fuel platforms.

The platforms that the translator imagines are means by which plain language queries can be translated into graph queries and have answers returned by some algorithmic "reasoning agent" that queries the Knowl-

improve clinical-trial design, and ultimately improve clinical care. This final assertion represents the driving motivation for the Translator system. [90]

[16] In an odd mixture of metaphors, members of the Translator consortium introduced the project with a piece titled "Deconstructing the Translational Tower of Babel." [91] A common interpretation of the Biblical Tower of Babel is as a symbol of the hubris of humanity, attempting to inscribe ourselves into the heavens and become immune to any future God-induced flooding. God, concerned with the power of a humanity unified under a single language, punished them by scattering people into groups speaking mutually unintelligible languages so they would not complete the tower. It is unclear why an effort to create a universalizing ontology would then be deconstructing a tower of babel, as it was the power of a unified language that allowed it to be built. Perhaps in other interpretations the Tower is an obelisk that suppresses the reunification of language. But I digress.

[17] The title of the Biolink paper is "Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science" [92]

[18] General as opposed to an ontology like MONDO [94] that identifies specific diseases.

[19] or links, labeled edges.

[20] To their credit, the Translator project seems to have made some of the long-delayed tooling for declaring a schema in a more accessible syntax[^yaml] than RDFS/OWL and generating representations in multiple formats, from JSON-LD to pydantic models. The Biolink paper also mentions a "Node Normalization Service" for being able to resolve Linked Data entities from different vocabularies that have been declared to be the same thing, but at the time of writing development seems to have slowed

[21] In a 2018 presentation by one of Biolink's authors: "What NOT to use the biolink-model for: Raw data, Metadata about a dataset" with some caveat that the underlying metamodel might still be useful [95] .

edge Providers and synthesizes a response [98, 99, 100, 92, 101] . We are not intended to use the data from Knowledge Providers directly, as it is likely to be incomplete or conflicting. Instead, the imagined use is as a recommendation system for researchers to target their research or for doctors to render care.

Several pilot experiments have demonstrated combining some aggregated patient records with the broader knowledge graph in order to, eg. identify new risk markers for disease [97, 102, 103, 104] . These systems layer personal records underneath "general" biomedical information like drug interactions and biological processes and use the extended information from the graph to infer information both about the nature of the disease and the patient. A platform integrated with the UCSF electronic health record system that layers disaggregated clinical records under the general knowledge graph is already apparently in a state of mature development [105] .

It is only with the inclusion of patient records into the knowledge graph that it becomes possible to use in a clinical setting: for even basic queries like "which drugs treat this disease" one has to be aware of patient qualities like allergies and comorbid conditions. To know how to treat the generic diagnosis of "gender dysphoria," one needs to know which gender the patient is experiencing dysphoria about. The logic of knowledge graph isn't just hungry for *some* personal medical data, the promise of the knowledge graph is that more data **always** improves the computations performed on it[22].

Why might we be critical about the NIH funding a series of projects to unify biomedical and personal health data in some universalized, platformatized knowledge graph? In short: because it won't work and will inevitably be captured by the surveillance industry.

First, as with any machine-learning based system, the algorithm can only reflect the implicit structure of its creation, including the beliefs and values of its architects [106, 107] , its training data and accompanying bias [108] , and so on. The "mass of data" approach ML tools lend themselves to, in this case, querying hundreds of independently operated databases, makes dissecting the provenance of every entry from every data provider effectively impossible. For example, one of the providers, mydisease.info was more than happy to respond to a query for the outmoded definition of "transsexualism" as a disease [109] along with a list of genes and variants that supposedly "cause" it - see for yourself. At the time of the search, tracing the source of that entry first led to the disease ontology DOID:1234, which has an official IRI, but in this case was being served by a graph aggregator Ontobee (Archive Link), which in turn listed this unofficial github repository **maintained by a single person** as its source[23]. This is, presumably, the fragility and inconsistency in input data that the machine learning layer is intended to putty over.

If the graph encodes being transgender as a disease, it is not farfetched to imagine the ranking system attempting to "cure" it. In a seemingly prerelease version of the translator's query engine, ARAX, it does just that: in a query for entities with a `biolink:treats` link to gender dysphoria[24], it ranks the standard therapeutics [110, 111] Testosterone and Estradiol 6th and 10th of 11, respectively — behind a recommendation for Lithium (4th) and Pimozide (5th) due to an automated text scrape of two conversion therapy papers[25]. Queries to ARAX for treatments for gender identity disorder helpfully yielded "zinc" and "water," offering a paper from the translator group that describes automated drug recommendation as the only provenance [112] . A query for treatments for `DOID:1233` "transvestism" was predictably troubling. The ROBOKOP [113] query engine behaved similarly, answering [a query for genes associated with]({{ "/data/ROBOKOP_message.json" | relative_url }}) gender dysphoria with exclusively trivial or incorrect responses[26].

It is critically important to understand that with an algorithmic, graph-based precision medicine system like this **harm can occur even without intended malice.** The power of the graph model for precision medicine

[22]The answer to a question posed as an algorithmic problem is always more data: "These results suggest that if more EHR concepts were mapped to SPOKE, a significant improvement in the classifier could be achieved." [102]

[23]I submitted a pull request to remove it, and a full year later it was merged!

[24]To its credit, ARAX does transform the request for `DOID:10919` to `MONDO:0001153` - gender dysphoria.

[25]as well as a recommendation for "date allergenic extract" from a misinterpretation of "to date" in the abstract of a paper that reads "Cross-sex hormonal treatment (CHT) used for gender dysphoria (GD) could by itself affect well-being without the use of genital surgery; however, **to date,** there is a paucity of studies investigating the effects of CHT alone"

[26]ITSN2 was identified in an unrelated paper about attachment patterns, HSD17B3 and 5a-RD2 were incorrectly identified as HSD17B13 and DHRS11 from another paper, POMC and OPN1SW were sourced from two papers that don't mention them. Androgen receptors were also identified, which is probably true, but almost trivially so.

is precisely its ability to make use of the extended structure of the graph[27]. The "value added" by the personalized biomedical graph is being able to incorporate the patient's personal information like genetics, environment, and comorbidities into diagnosis and treatment. So, harmful information embedded within a graph — like transness being a disease in search of a cure — means the system either a) incorporates that harm into its outputs for seemingly unrelated queries or b) doesn't work. This explodes the risk surface for medically marginalized people to include the entire Translator ecosystem: from the violence historically encoded in mainstream medical practices and ontologies (eg. [109, 114] , among many), to incorrectly encoded information like that from automated text mining, to explicitly adversarial information injected into the graph through some crowdsourcing portal like this one [115] . Each of these sources of harm could influence medical care in ways that *even a well-meaning clinician might not be able to recognize.*

The risk of harm is again multiplied by the potential for harmful outputs of a biomedical knowledge graph system to trickle through medical practice and re-enter as training data. The Consortium also describes the potential for ranking algorithms to be continuously updated based on usage or results in research or clinical practice[28] [90] . Existing harm in medical practice, amplified by any induced by the Translator system, could then be re-encoded as implicit medical consensus in an opaque recommendation algorithm. There is, of course, no unique "loss function" to evaluate health. One belief system's vision of health is demonic pathology in another. Say an insurance company uses the clinical recommendations of some algorithm built off the Translator's graph to evaluate its coverage of medical procedures. This gives them license to lower their bottom line under cover of some seemingly objective but fundamentally unaccountable algorithm. Could a collection of anti-abortion clinics giving one star to abortion in every case meaningfully influence whether abortion is prescribed or covered? Why not? Who moderates the graph?

The centralized structure of the Translator's Knowledge Providers and query engines now form a small group responsible for curating the entire structure of biomedical information. The curation process could be "crowdsourced" to allow affected communities to suggest improvements, but the platformatized nature of the Translator both concentrates decisionmaking power and diffuses responsibility across a string of platform holders. Who is supposed to fix incorrect or harmful query responses? Is it the responsibility of the potentially dozens of Knowledge Providers, the swarm of reasoning agents, or the frontend wrapper you pay a monthly subscription for? It is the platformatized nature of the Translator itself that creates the need for centralized moderation in the first place. The design of the Translator to evolve into a series of "user-" or customer-facing platforms that aspire to universality binds it to all the regulatory burden any biomedical technology bears. The cost of moderation will of course be enormous, placing a fundamental constraint on its lifespan as a publicly funded project — and a strong incentive towards co-option by the information conglomerates capable of paying it[29].

These problems hint at the likely fate of the Translator project. Rather than integrating into the daily practice of researchers, the centralized process of creating Knowledge Providers can only be maintained for as long as the grant funding for the Translator project lasts. When queried at the time of writing, of the 25 knowledge providers that were responsive to information about "Anything that is related to the common cold," 22 were unresponsive or timed out. How the Translator is intended to work by its architects is almost irrelevant compared to the question of what happens to it *after the project ends.*

[27] eg. Some members of the SPOKE project, a Knowledge Provider for the Translator project, describe the effects of the extended graph as "pushing" or influencing the "flow" of information: > "For this patient, information flows from Carbamazepine to a set of Disease nodes (either through "treated by" or "contraindicated for" edges) and then (either directly or through an additional Disease or Gene node) to the genes CNP, MAG, or PTEN which are all components of "Myelin sheath adaxonal region." [102]

[28]

> "The Reasoners then return ranked and scored potential translations with provenance and supporting evidence. The user is then able to evaluate the translations and supporting evidence and provide feedback to the Reasoners, thus promoting continuous improvement of the prototype system." [90]

[29] There is a clear analogy to the recent push to increase internet content regulation by social media companies [116] . A platform makes a quasi-universal social space for profit, moderation then has to scale with the size of the platform, then it lobbies to increase regulatory burden to a point that is impossible to maintain for all but already-scaled companies. It is only the quasi-universality of the platform that makes the moderation burden so high in the first place, however, compared to eg. a decentralized medium that might have a structurally different disposition to moderation (see [117]).

Linking biomedical and patient data in a single platform is a natural route towards a multisided market where records management apps are sold to patients, treatment recommendation systems are sold to clinicians, research tools and advertising opportunities are sold to pharmaceutical companies, risk metrics are sold to insurance companies, and so on. The contours of this market are already clear.

As a non-exhaustive set of examples:

- I have already described **RELX**'s interest in personal biomedical data. Their 2022 Annual Report [37] is the first year where they explicitly describe their entrance into the patient data market[30]. RELX is a particularly worrying example because of their established roles among academics, medical workers, and insurance providers.
- **Amazon** already has a broad home surveillance portfolio [118] , and has been aggressively expanding into health technology [119] and even literally providing health care [120, 121] , which could be particularly dangerous with the uploading of all scientific and medical data onto AWS with entirely unenforceable promises of data privacy through NIH's STRIDES program [122] .
- **Google** already includes medical conditions in its surveillance-backed advertising profiles [123, 124] , and is edging its way into wearable health data with eg. its acquisition of FitBit [125] . It also already has a system, Med-PALM, for biomedical question answering based on large language models [126, 127, 128] . Search is a primary entrypoint for many people searching for health information, and Google presumably would be more than happy to merge that data with a generalized biomedical knowledge graph.
- **Apple** already has a matured Health ecosystem of apps and services for both patients, clinicians, and researchers [129, 130] and has a similar exposure to relevant data and control of platforms (iOS, watchOS) to make use of it, though they have marketed themselves in the surveillance space as a defender of privacy.
- Of course **Microsoft** [131] and **IBM** [132] are also in play.

The design of the Translator project reflects the prevailing logic of the surveillance economy as powered by knowledge graphs, and is poised to be swallowed up by it. Rather than a means for us to collectively make sense together, they have imagined a cloud-driven system where a small group of experts wave a wand of unknowable algorithms over a bulging plastic trash bag of data to pull out the Magic Knowledge Rabbit. The noble intention of making a generalized biomedical knowledge graph for the public good is unlikely to be realized. In the process, though, the NIH will have funded facilitating technologies and standards for the merger of personal medical surveillance with the broader landscape of biomedical data. Academics will have new vectors by which they become unwitting or unwilling collaborators with surveillance and data brokers, lending what credibility they have left to a landscape of buggy black boxes of biopolitical control. And, most importantly, vulnerable populations will have dozens of new ways to be marginalized by the techno-political medical establishment.

3.0.3 NSF: Open Knowledge Network

While the NIH builds a set of universal knowledge graphs for biomedical information, the NSF is building them for everything else. Its Open Knowledge Network (OKN) project intends to "provide an essential public-data infrastructure for enabling an AI-driven future." [133] OKN is in an earlier stage of development than the Translator, so this section is less focused on the details of individual projects and more to argue the pattern of public/private knowledge graphs is an emerging consensus.

Compared to the Translator, the OKN pulls punches for neither its utopian promises nor obvious risks. Some sections of its roadmap are written in a style where each line shoots for the stars because even if some of

[30] In commercial healthcare, identity, claims and provider data is combined with patient information to assist healthcare providers, pharmacies and insurers in delivering improved health outcomes, ensuring accurate and complete provider data and regulatory compliance. [37]

them miss the result is a constellation of absolute bangers like "Harnessing the vast amounts of data generated in every sphere of life and transforming them into useful, actionable information and knowledge is crucial to the efficient functioning of a modern society[31]" [133] . The project was initially proposed in 2017, went through two cohorts of projects within the NSF Convergence Accelerator in 2019 and 2020[32], and invited a broader submission of proposals in November 2021 [135] . The roadmap comes at the end of a series of workshops in 2022 intended to scope and outline the OKN, so there is still very little public evidence of its progress to evaluate[33].

Its domain is much broader than the Translator, and is unmistakeably bound up in both the United States Federal Government's military and political interests in Artificial Intelligence [139] and the information economy's interests in making a universal space where all information can be bought and sold with minimal friction [140] . Where the Translator has the near-inevitable risk of being captured by information conglomerates, through the euphemism of "public private partnership" the OKN makes clear it was already captured at inception: the team behind the SPOKE biomedical knowledge network immediately spun off a for-profit startup to sell the graph as a cloud service [141] , abandoning further UX development of its publicly accessible demo.

Without mincing words, the OKN intends to make a Universal Knowledge Graph of Everything. They check all the boxes[34]: a) make authoritative schemas for everything, b) link them all together, c) ingest data from as many sources as possible at whatever quality available, d) integrate private with public data e) put it all in the cloud! (p. 18-19 "Creating an OKN" [140]).

They OKN describes its work using a vocabulary of "vertical" "horizontal," where "vertical" applications refer to specific uses or domains like energy or health data, and "horizontal" themes like technologies and governance are shared across all domains. The work of the OKN is organized around specific use cases either within a "vertical" topic or a specific "horizontal" theme with the intent of later building them together into a shared infrastructure. The collection of "vertical" topics identified in the 2022 roadmap hint at the effectively unbounded scope of the OKN: accelerated capitalism via supply chain logistics, more tightly integrated weapons development, a handful of climate change projects, an omniscient financial system, and so on. Each imagines the primary problem in a given domain not as structural exploitation or injustice, but a lack of data[35].

A collection of "vertical" topical working groups in the 2022 roadmap centered on an algorithmic justice system are illustrative: An **Integrated Justice Platform** group describes how greater surveillance across every contact people have with the US Justice System is necessary to decrease bias. The group outlines a wish list of data sources they would integrate - arrest and booking, jail, trial, prosecution, and the rest. A **Decarceration** group[36] describes extending that surveillance through to the rest of incarcerated people's lives after they are released - rehab, parole, foster care, shelters, public services, etc. A **Homelessness** group intends to track unhoused people in order to match them to available resources. A **Decision Support for Government** group describes bundling up these and other data sources into platforms for making "data driven decisions" on topics including crime and policing.

On their own, each of these groups describes noble goals: decreasing bias in the justice system, providing resources to formerly incarcerated or unhoused people, making government decisions more efficient. Taken together, however, the projects describe a panoptical surveillance system that wouldn't even need to be reconfigured to be used for algorithmically-enhanced oppression. I doubt any of the researchers in these groups intend for their work to be used for state violence, but *Palantir doesn't care what academics intended their tools to be used for*[37].

[31] If you could rig an MS Word template to punctuate sentences with "Whoomp! (There It Is)," they would have.

[32] The Convergence Accelerator is a project specifically designed to provide public research funding to for-profit industries [134]

[33] SPOKE, discussed previously, was funded by both the Translator project [136] and OKN [137] , and KnowWhereGraph is another notable early prototype [138]

[34] Hopefully this pattern is familiar.

[35] A recurring pattern in techno-solutionism: > "These perspectives assume that complex controversies can be solved by getting correct information where it needs to go as efficiently as possible. In this model, political conflict arises primarily from a lack of information. If we just gather all the facts, systems engineers assume, the correct answers to intractable policy problems like homelessness will be simple, uncontroversial, and widely shared. > > But, for better or worse, **this is not how politics work.**" [142]

[36] Including a representative from Booz Allen Hamilton, which may be familiar as the former employer of Edward Snowden, who was working for them on a contract with the NSA which gave him access to the details of its PRISM mass-surveillance program.

[37]

The motivations behind integrating government data sources and automating public benefit delivery cannot overcome the context of systemic oppression they are embedded within. Group H, the "Homelessness OKN" group, takes particular effort[38] to focus on the needs of the unhoused and address the potential risks of "track[ing] homelessness in real time, [and] identify[ing] available homelessness programs and services," but misses the already-real harms of similar prior efforts. Virginia Eubanks describes how Los Angeles County's Coordinated Entry System — a program very much like that described by group H, intended to match unhoused people with housing supply by integrating previously siloed data systems — operates as a sophisticated mechanism of control and punishment:

> For Gary Boatwright and tens of thousands of others who have not been matched with any services, coordinated entry seems to collect increasingly sensitive, intrusive data to track their movements and behavior, but doesn't offer anything in return. [...] Moreover, the pattern of increased data collection, sharing, and surveillance reinforces the criminalization of the unhoused, if only because **so many of the basic conditions of being homeless are also officially crimes.** [...] The tickets turn into warrants, and then law enforcement has further reason to search the databases to find "fugitives." Thus, **data collection, storage, and sharing in homeless service programs are often starting points in a process that criminalizes the poor.** [...]
>
> Further integrating programs aimed at providing economic security and those focused on crime control threatens to turn routine survival strategies of those living in extreme poverty into crimes. **The constant data collection from a vast array of high-tech tools wielded by homeless services, business improvement districts, and law enforcement create what Skid Row residents perceive as a net of constraint that influences their every decision.** Daily, they feel encouraged to self-deport or self-imprison. Those living outdoors in encampments feel pressured to constantly be on the move. Those housed in SROs or permanent supportive housing feel equally intense pressure to stay inside and out of the public eye. [...] **Coordinated entry is not just a system for managing information or matching demand to supply. It is a surveillance system for sorting and criminalizing the poor.** [142]

It is impossible to consider integrated data in government without confronting the reality of algorithmic policing. Under its Strategic Plan goal of "Realiz[ing] Tomorrow's Government Today" Los Angeles County has already been integrating its information systems, including creating a unified system of law enforcement and other public service data "to identify super utilizers of justice and health system resources"[39] [145, 146] . Many police departments — including the LAPD — already have access to the kind of linked data ecosystems described by the OKN by renting them from private data brokers like Palantir [143, 147] . These data infrastructures facilitate the well-described feedback loop of predictive policing, where areas already subject to historical economic and racist violence are classified as "high-crime areas," more police are concentrated there, in turn causing them to measure or create more crime[40] [143, 148, 150, 149, 151, 152, 153] . The reformist idea that more data will help us "police the police" is belied by the resolute history of more data allowing the police to innovate on information asymmetries to create new expressions of power [154, 155] .

> "Because one of Palantir's biggest selling points is the ease with which new, external data sources can be incorporated into the platform, its coverage grows every day. LAPD data, data collected by other government agencies, and external data, including privately collected data accessed through licensing agreements with data brokers, are among at least 19 databases feeding Palantir at JRIC." [143]

[38] Given the context of the "Innovation Sprint" essentially as an extended pitch session for future work, and the disincentive towards serious discussion of the ethical ramifications of the projects that entails.

[39] ...and then outsourcing the maintenance and risk of it being breached [144]

[40]

> These visits often resulted in other, unrelated arrests that further victimized families and added to the likelihood that they would be visited and harassed again. In one incident, the mother of a targeted teenager was issued a $2,500 fine when police sent to check in on her child saw chickens in the backyard. In another incident, a father was arrested when police looked through the window of the house and saw a 17-year-old smoking a cigarette. These are the kinds of usually unreported crimes that occur in all neighborhoods, across all economic strata—but which only those marginalized people who live under near constant policing are penalized for. [148, 149]

The critical difference between prior infrastructures and those imagined by the OKN is that they are explicitly designed to be linked into a continuous network of data that enables the same kind of data-driven decisionmaking that drives predictive policing for *any* system. We should not be imagining the utterly mechanistic bureaucracy of *Kafka* here, but rather the deeply expressive and personal exercise of power of Terry Gilliam's *Brazil*. Widespread algorithmic governance doesn't necesarily look like a faceless bureaucracy where all decisions are made by a computer, existing algorithmic systems like predictive policing and the working conditions at Amazon warehouses retain the very human domain of *discretion* (see [154]). The algorithms and seemingly open infrastructures purport themselves as objective and egalitarian, but who they are built for, who gets to provides the inputs, and who decides which outputs matter make their reality very different. The very act of creating information infrastructures intended to algorithmically solve the world's problems is itself an expression of power-based discretion that diffuses energy that might be better spent elsewhere: rather than attempt to address the root cause, we can make a big show of *doing something* by diverting a large amount of resources and labor to gathering data and deriving "insights" about them.

The prominent role of climate change among the topics identified by the OKN project is at once reassuring and damning. Maybe what we need to solve climate change isn't data, it's to organize effective climate movements outside of the algorithmically disorienting information ecosystems designed to disorient us with engagement-maximizing rage-bait and transmute all movement building into influencer culture. Maybe what we need to address mass poverty isn't data, it's to dismantle the mechanisms of mass extraction that are increasingly powered by economies of surveillance. Maybe what we need to make the criminal justice system less racist isn't more data to feed into their predictive policing algorithms, but to abolish the police.

In both of these projects, the pursuit of a universal knowledge graph is motivated by altruistic goals — that those goals take the form of the universal knowledge graph is not a matter of stupidity or malice, but ideology.

4 Infrastructural Ideologies

The emerging models

- the critical thing here is that the US's two major public funding bodies have lined up to build unifying graphs of everything for ostensibly noble purposes but the information companies are literally licking their chops to eat this up. They have both effectively made new funding mechanisms that allow them to basically let private industry types what to build and how to build it, the academics get invited in to play for a bit, and then it's time to capture it later. The specifics aren't important, what's important is the pattern of layering public and private knowledge graphs that seem great and open on the surface but actually power some monstrously shittier technology underneath.
- Census of outcomes at the end - it could be really bad, it could actually not amount to anything at all, but regardless this is illustrative of the way that the cloud constrains our thinking! Either nonfunctional or corporate! Note how much has been spent on each and give caveats like "ya these might not even amount to anything, but the point is the way that the existing cloud paradigm constrains the way that we think precisely in such a way that it is difficult to actually build the kind of effective infrastructure that we want. We have to acknowledge that the cloud model is not a universal computing paradigm, but one designed for a very specific mode of property relationships and profit models."

Scraps:

- I mean what could be different? how do we get out of the loop of trying to make "ok the real unified knowledge base for real this time guys" loop [156] . Maybe it's the belief that there should be a single unified naming system that's pathological? Maybe we need to focus on linking these things together rather than projecting them into some singular space.

- it is, in fact ideological! "SPOKE was conceived with the philosophy that if relevant information is connected, it can result in the emergence of knowledge, and hence provide insights into the understanding of diseases, discovering of drugs and proactively improving personal health." [97]
- the hollow middle, right? like why is it so hard for me to even find all the ontologies that have been made by the biomedical translator project? why are most of the API endpoints in arax dead *already* - it hasn't even launched! why aren't there the tools for me to actually *use the information* in a way that wasn't specifically designed for in the platform? by focusng the design into these API-based data sources with platforms strapped on top of them, the project is limited to *only exactly what the developers could imagine and are capable of doing* rather than fulfullig the initial definition (find cite) of what "platforms" meant - an enabling technology that allows an expansive space of use that goes beyond the creators and enables and empowers people!

4.1 Surveillance Graphs

- read and cite:
  - [69]

Why on *earth* would we want our insurance provider to be able to adjust their premiums based on our private medical information, even if it is "de-identified"

The problem of belief: what should these things do? present us instantly with results? or should they be part of a messy dialectic of information where the graph itself is the outcome? The cloud orthodoxy suggests the former!

- Need to articulate the components of the cloud orthodoxy mindset
  - Automated reasoning
  - More data is better
  - Neutral reflection of underlying facts, only conditioning factor is 'confidence', but no criticism of schema
  - End goal is productized services: eg. the fantasy of the google researcher in 2018 at the end of entity oriented search
  - Convenience
  - Performance
  - Assumes that we are all subjects of surveillance, that the data is "out there" to be mined, and its heterogeneity is an inconvenience or technical problem.
  - There is "one" of something - there is one index that refers to the concept of a man, or a given person, etc. rather than many contextual representations of that thing.
    - the poltics of ontologies are neatly illustrated by the fact that "a woman is a woman" means precisely the opposite thing w.r.t trans woman to transmisics as compared to normal people.
  - It doesn't have to work! as long as it seems like it does
  - Rather than building tools for people that they might use to actually integrate the system in their work or lives, but instead some means of scraping it afterwards (eg. SCALES), so who is this actually for? the people studying people doing work, or for the people doing the work? When things are intended for the people actually doing the work, they are invariably platforms that provide just enough functionality to keep people using them while surveilling them as the main purpose.
- YOU ONLY NEED TO DO THIS SCALE OF AUTOMATED EXTRACTION IF YOU PRESUPPOSE A UNIVERSE OF A FEW MASSIVE PLATFORMS OWNING EVERYTHING IN THE FIRST PLACE re: amazon product graph example in [2]

- The argument that the next phase of everything is chatbots and

The merger of AI shit and knowledge graph shit

So if web 2.0 was about the platformatization of the web, what is up next? In the same way that information conglomerates successfully harnessed open source for profit, now can we effectively gamify the entirety of knowledge generation process? can we put ppl inside of information curation chambers when they search?

"Explainable AI" is gonna come from knowledge graphs yo

Broad principles - Strategic use of "openness" when it facilitates greater market control and to prevent someone else from capturing a particular element of the technology. - Same thing as with the DOI: the linking agreements were killing us! Get some barrier to commerce out of the way (ontology discontinuity) so that the commerce can intensify, not so that it can abate

> Peter Mika: A natural next step for Knowledge Graphs is to extend beyond the boundaries of organisations, connecting data assets of companies along business value chains. This process is still at an early stage, and there is a need for trade associations or industry-specific standards organisations to step in, especially when it comes to developing shared entity identifier schemes. [88]

4.1.1 Knowledge Graphs + AI

MOVE THIS to "models/surveillance" - the cognitive style of search, why chatbots inevitable and an intrinsic part of the plan, and the broader conversation of the surveillance mindset

Scraps: - Role of knowledge graphs in explainable AI: [157] - Explicit algo + KG model - entity oriented search: [158] - Microsoft re: academic graph with chat-based search: [131] - cognitive expectations of search - 2002 RDF Core WG talking about whether we want things to work like google or not... [159] - https://lists.w3.org/Archives/Public/w3c-rdfcore-wg/2002Sep/0276.html - agents vs. platforms: - where are all the agents? we were supposed to have web robots by now!? [160] - they're doign business stuff! [161] - OKN is Specifically trying to get the chatbots up to speed: [140] - Microsoft's very fun assistant: [162]

Tim BL actually really wanted to be able to talk to the computers, he specifically imagined chatbots working on top of the semantic web: https://www.w3.org/DesignIssues/Evolution.html

The 'one bar' search paradigm is powerful and inculcates a very specific expectation of use... (expectations from search results)

the problem is that it naturally sacrifices all the extra query structure inherent in any "advanced search" interface.

But even an "advanced search" interface is the wrong metaphor, because (depending) those still give the expectation that, were you to parameterize your search correctly, you would receive a list of all matching results. So even if ordered by some anonymous "relevance" parameter, one could see "all" of something.

One-bar search sacrifices more than that and gives an expectation that getting exhaustive answers is not possible, no matter how hard one were to search, and that 'best enough' with some ranking is the best one can ask for. The ranking then takes on a different character, rather than ordering some finite list, it defines the contents of the list. So algorithmic search as we know it.

That relates to the initial goal of semantic web, to be able to give additional parameters/handles/structuring information to the web so that it was possible to do those kinds of "advanced searches" on your own, without needing a search engine.

What is lost in the single bar search has to be made up for in some way, so that falls on the ability to parse semantic meaning in the search query, as well as inject context from surveillance data.

Chatbots then are a means of expanding that context, specifically in such a way that the "local neighborhood" of some decision tree were being presented to you in plain language. But rather than, again, an indexical cognitive pattern where you expect to see everything, your being constrained to a particular 'neighborhood' of the graph space and how you are steered then becomes the product.

So chatbot search is a very natural match for the knowledge graphs that already parameterize the search and compute semantic meaning in the query.

Thus the criticism that LLM's don't "know anything" won't be true for long - they will be part of a joint system that decodes the search query, constructs a context, and then queries databases of structured data. (cite google paper to this effect)

Think broader than search engines though, the pernicious and dangerous part here is that we could merge several classes of platform and surveillance harm: individual surveillance could merge with medical and public information and insurance information and the rest in an interoperable interchange format so that the data brokering economy woudl effectively explode. Imagine the splintering of infinitely many platforms that each owned some subset of the data, each platform holder owning all of them and slicing them off to you and pocketing the costs.

merged with research and reference data, they could literally make a graph of all information and supplant libraries, etc. for all information from news to government to personal.

- KGs + AI are gonna try and do "explainable AI" - [163, 164]
  - Microsoft: KGs necessary for chatbots - "the LLMs don't *know anything* [165] . Google: [164]
  - AI is the future of KGs - [2]

  "Second, there is a clear recognition that KG representations are a central ingredient to achieving the compositional behavior in AI systems." - They want to integrate these with your personal knowledge graph: Google alredy owns all your calendar/etc. data, now they want you to take notes in a way that can be more easily mined: [166] . Under the guise of convenience - "update my location across all my apps"

- The evolution into chatbots
  - These companies want to be able to provide some interface to this graph without revealing it, chatbots are effectively a way to launder knowledge graphs into some information-consumer facing interface that lets them traverse the graph with natural language. Eg. you can hold the context of the graph in mind when doing subsequent searches
  - They want to make the processes of curating their graphs more interactive! A new kind of information serfdom - presumably google wants you to tell it when its information box is incorrect!?
  - Relationship between graphs and surveillance technologies: [52]

- And the broader universe of surveillance
  - Policing: [167, 143]
  - NSF graph describes military applications by the boatload

- The fundamental contradiction of platform capitalism

4.2 Vulgar Linked Data

vulgar LD is when we own it, rather than merely annotate things to be owned by google. what is the point where the owner of the graph can just say... no?

References

[1] McKenzie Wark. *Capital Is Dead: Is This Something Worse?* Verso Books, February 2021. ISBN 978-1-78873-533-9. 1, 2.0.3

[2] Vinay K. Chaudhri, Chaitanya Baru, Naren Chittar, Xin Luna Dong, Michael Genesereth, James Hendler, Aditya Kalyanpur, Douglas B. Lenat, Juan Sequeda, Denny Vrandečić, and Kuansan Wang. Knowledge graphs: Introduction, history, and perspectives. *AI Magazine*, 43(1):17–29, 2022. ISSN 2371-9621. https://doi.org/10.1002/aaai.12033. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aaai.12033. 2, 2.0.3, 4.1, 4.1.1

[3] Pascal Hitzler. A review of the semantic web field. *Communications of the ACM*, 64(2):76–83, January 2021. ISSN 0001-0782. https://doi.org/10.1145/3397512. URL https://doi.org/10.1145/3397512. 2, 2.0.1, 2.0.2

[4] Jihong Yan, Chengyu Wang, Wenliang Cheng, Ming Gao, and Aoying Zhou. A retrospective of knowledge graphs. *Frontiers of Computer Science*, 12(1):55–74, February 2018. ISSN 2095-2236. https://doi.org/10.1007/s11704-016-5228-9. URL https://doi.org/10.1007/s11704-016-5228-9. 2

[5] Mike Bergman. A Common Sense View of Knowledge Graphs, July 2019. URL https://www.mkbergman.com/2244/a-common-sense-view-of-knowledge-graphs/. 2

[6] Lisa Ehrlinger and Wolfram Wöß. Towards a Definition of Knowledge Graphs. *Semantics*, September 2016. 2

[7] Jonny L. Saunders. Decentralized Infrastructure for (Neuro)science, August 2022. URL http://arxiv.org/abs/2209.07493. 2, 3.0.1, 10, 3.0.2

[8] Tim Berners-Lee. Links and Law, April 1997. URL https://www.w3.org/DesignIssues/LinkLaw. 2.0.1

[9] Tim Berners-Lee. Links and Law: Myths, April 1997. URL https://www.w3.org/DesignIssues/LinkMyths.html. 2.0.1

[10] Tim Berners-Lee. What the Semantic Web can Represent, September 1998. URL https://www.w3.org/DesignIssues/RDFnot.html. 2.0.1, 2.0.3

[11] Tim Berners-Lee. The Scale-free nature of the Web, 1998. URL https://www.w3.org/DesignIssues/Fractal.html. 2.0.1

[12] Tim Berners-Lee, James HENDLER, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, 2001. ISSN 0036-8733. URL https://www.jstor.org/stable/26059207. 2.0.1

[13] Tim Berners-Lee. Cultures and Boundaries, July 2007. URL https://www.w3.org/DesignIssues/Culture.html. 2.0.1

[14] Sean B. Palmer. Ditching the Semantic Web?, March 2008. URL http://inamidst.com/whits/2008/ditching. 2.0.2

[15] Aaron Swartz. Aaron Swartz's A Programmable Web: An Unfinished Work. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 3(2):1–64, February 2013. ISSN 2160-4711, 2160-472X. https://doi.org/10.2200/S00481ED1V01Y201302WBE005. URL http://www.morganclaypool.com/doi/abs/10.2200/S00481ED1V01Y201302WBE005. 2.0.2

[16] Lindsay Poirier. A Turn for the Scruffy: An Ethnographic Study of Semantic Web Architecture. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, pages 359–367, New York, NY, USA, June 2017. Association for Computing Machinery. ISBN 978-1-4503-4896-6. https://doi.org/10.1145/3091478.3091505. URL https://doi.org/10.1145/3091478.3091505. 2.0.2

[17] Tim Berners-Lee. Linked Data, July 2006. URL https://www.w3.org/DesignIssues/LinkedData.html. 2.0.2, 2.0.3

[18] Amit Singhal. Introducing the Knowledge Graph: Things, not strings, May 2012. URL https://blog.google/products/search/introducing-knowledge-graph-things-not/. 2.0.3

[19] Iain. Freebase is dead, long live Freebase, May 2016. URL https://medium.com/@iainsproat/freebase-is-dead-long-live-freebase-6c1daff44d19. 2.0.3

[20] Natasha Noy, Yuqing Gao, Anshu Jain, Anant Narayanan, Alan Patterson, and Jamie Taylor. Industry-scale Knowledge Graphs: Lessons and Challenges: Five diverse technology companies show how it's done. *Queue*, 17(2):Pages 20:48–Pages 20:75, April 2019. ISSN 1542-7730. https://doi.org/10.1145/3329781.3332266. URL https://doi.org/10.1145/3329781.3332266. 2.0.3, 3, 3.0.1

[21] Ben Tarnoff. *Internet for the People: The Fight for Our Digital Future*. Verso Books, June 2022. ISBN 978-1-83976-202-4. 2.0.3

[22] Shoshana Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books, London, 2019. ISBN 978-1-78125-685-5. 2.0.3

[23] Tim Berners-Lee. Socially aware cloud storage, August 2009. URL https://www.w3.org/DesignIssues/CloudStorage.html. 2.0.3

[24] Tim Berners-Lee. Goals for a Human-Data Interface, July 2010. URL https://www.w3.org/DesignIssues/TabulatorGoals.html. 2.0.3

[25] Neo4j. Neo4j Customers. URL https://neo4j.com/customers/. 2.0.3

[26] Enterprise Knowledge Graph Foundation and Michael Atkin. Knowledge Graph Industry Survey Report, October 2022. URL https://www.ontotext.com/knowledgehub/white_paper/knowledge-graph-industry-survey-report/. 2.0.3

[27] Jennifer L. Schenker. New Report Details Industry's Use of Knowledge Graphs, May 2021. URL https://theinnovator.news/new-report-details-industrys-use-of-knowledge-graphs/. 2.0.3, 5

[28] Juan Sequeda and Ora Lassila. Designing and Building Enterprise Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge, Cham, 2021. Springer International Publishing. ISBN 978-3-031-00788-0 978-3-031-01916-6. https://doi.org/10.1007/978-3-031-01916-6. URL https://link.springer.com/10.1007/978-3-031-01916-6. 2.0.3

[29] Antonia Azzini, Sylvio Barbon, Valerio Bellandi, Tiziana Catarci, Paolo Ceravolo, Philippe Cudré-Mauroux, Samira Maghool, Jaroslav Pokorny, Monica Scannapieco, Florence Sedes, Gabriel Marques Tavares, and Robert Wrembel. Advances in Data Management in the Big Data Era. In Michael Goedicke, Erich Neuhold, and Kai Rannenberg, editors, *Advancing Research in Information and Communication Technology*, volume 600, pages 99–126, Cham, 2021. Springer International Publishing. ISBN 978-3-030-81700-8 978-3-030-81701-5. https://doi.org/10.1007/978-3-030-81701-5_4. URL https://link.springer.com/10.1007/978-3-030-81701-5_4. 2.0.3

[30] Suresh Toby Segaran. Two-phase construction of data graphs from disparate inputs, April 2020. URL https://patents.google.com/patent/US10614127B2/. 2.0.3

[31] Paolo Ceravolo, Antonia Azzini, Marco Angelini, Tiziana Catarci, Philippe Cudré-Mauroux, Ernesto Damiani, Alexandra Mazak, Maurice Van Keulen, Mustafa Jarrar, Giuseppe Santucci, Kai-Uwe Sattler, Monica Scannapieco, Manuel Wimmer, Robert Wrembel, and Fadi Zaraket. Big Data Semantics. *Journal on Data Semantics*, 7(2):65–85, June 2018. ISSN 1861-2032, 1861-2040. https://doi.org/10.1007/s13740-018-0086-2. URL http://link.springer.com/10.1007/s13740-018-0086-2. 2.0.3

[32] Maya Natarajan. From Graph To Knowledge Graph: A short journey to unlimited insights. URL https://neo4j.com/whitepapers/knowledge-graphs-unlimited-insights/thanks/. 2.0.3

[33] Dean Allemang. Merging data graphs made easy, December 2022. URL https://scribe.rip/@dallemang/merging-data-graphs-made-easy-8b7e616acfe6. 2.0.3

[34] Dean Allemang. Merging tables is hard, December 2022. URL https://scribe.citizen4.eu/@dallemang/merging-tables-is-hard-89d8637a081. 2.0.3

[35] Boris Villazon-Terrazas, Nuria Garcia-Santa, Yuan Ren, Alessandro Faraotti, Honghan Wu, Yuting Zhao, Guido Vetere, and Jeff Z. Pan. Knowledge Graph Foundations. In Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu, editors, *Exploiting Linked Data and Knowledge Graphs in Large Organisations*, pages 17–55. Springer International Publishing, Cham, 2017. ISBN 978-3-319-45654-6. https://doi.org/10.1007/978-3-319-45654-6_2. URL https://doi.org/10.1007/978-3-319-45654-6_2. 2.0.3

[36] Sarah Lamdan. *Data Cartels: The Companies That Control and Monopolize Our Information*. Stanford University Press, Stanford, California, 2023. ISBN 978-1-5036-1507-6 978-1-5036-3371-1. 2.0.3

[37] RELX. Annual Report 2022, February 2023. URL https://www.relx.com/~/media/Files/R/RELX-Group/documents/reports/annual-reports/relx-2022-annual-report.pdf. 2.0.3, 3.0.2, 30

[38] Roger C. Schonfeld. A Reorganization at Elsevier, May 2022. URL https://scholarlykitchen.sspnet.org/2022/05/02/reorganization-elsevier/. 7

[39] Elsevier and EMD Serono. Making medical information easily accessible to healthcare professionals, 2021. URL https://www.elsevier.com/__data/assets/pdf_file/0004/1276087/Elsevier-EMD-Serono-phactMI-collaboration.pdf. 2.0.3

[40] Elsevier. Rethink Clincal Content, April 2020. URL https://www.elsmediakits.com/intelligence/white-paper/white-paper%3A-rethink-clinical-content. 2.0.3

[41] Sam Biddle. LexisNexis to Provide Giant Database of Personal Information to ICE, April 2021. URL https://theintercept.com/2021/04/02/ice-database-surveillance-lexisnexis/. 2.0.3

[42] Sam Biddle. ICE Searched LexisNexis Database Over 1 Million Times in Just Seven Months. *The Intercept*, June 2022. URL https://theintercept.com/2022/06/09/ice-lexisnexis-mass-surveillances/. 2.0.3

[43] LexisNexis Risk Solutions. Accurint® TraX™, . URL https://risk.lexisnexis.com/products/accurint-trax. 2.0.3

[44] LexisNexis Risk Solutions. Telematics OnDemand, . URL https://risk.lexisnexis.com/products/telematics-ondemand. 2.0.3

[45] LexisNexis Risk Solutions. Accurint for Legal Professionals, April 2022. URL https://web.archive.org/web/20230308034302/https://www.lexisnexis.com/pdf/AccurintForLegalProfessionals/24.pdf. 2.0.3

[46] LexisNexis Risk Solutions. LexID, . URL https://risk.lexisnexis.com/our-technology/lexid. 2.0.3

[47] Neo4j. Neo4j + U.S. Army Case Study, 2021. URL https://neo4j.com/case-studies/us-army/. 2.0.3

[48] David Cohen, Kevin Richards, and Khan Tasinga. System and Method for Sharing Investigation Result Data, November 2015. URL https://patents.google.com/patent/AU2013251186B2/. 2.0.3

[49] Shivam Mathura, Lucas Lemanowicz, and Tim Vergenz. Automated database analysis to detect malfeasance, August 2017. URL https://patents.google.com/patent/US20170221063A1/. 2.0.3

[50] Timothy Yousaf, Alexander Mark, Sharon Hao, David Cohen, Andrew Elder, Daniel Lidor, Joel Ossher, Christopher RICHBOURG, Joshua Zavilla, and Kevin Zhang. Systems, methods, user interfaces and algorithms for performing database analysis and search of information involving structured and/or semi-structured data, January 2018. URL https://patents.google.com/patent/US9881066B1/. 2.0.3

[51] Eric Knudson, Matthew Gerhardt, Andrew Elder, and Eli Rosofsky. Systems and methods for annotating and linking electronic documents, January 2021. URL https://patents.google.com/patent/US20210004530A1/. 2.0.3

[52] Andrew Iliadis and Amelia Acker. The seer and the seen: Surveying Palantir's surveillance platform. *The Information Society*, 38(5):334–363, October 2022. ISSN 0197-2243, 1087-6537. https://doi.org/10.1080/01972243.2022.2100851. URL https://www.tandfonline.com/doi/full/10.1080/01972243.2022.2100851. 2.0.3, 4.1.1

[53] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. MIT Press, Cambridge, MA, USA, September 1999. ISBN 978-0-262-02461-7. 3.0.1, 3.0.2

[54] Wendy Liu. Freedom Isn't Free, August 2018. URL https://logicmag.io/failure/freedom-isnt-free/. 3.0.1

[55] McKenzie Wark. *A Hacker Manifesto*. Harvard University Press, October 2004. ISBN 978-0-674-01543-2. URL https://www.hup.harvard.edu/catalog.php?isbn=9780674015432. 3.0.1

[56] Daniel Goldsmith. The Original Sin of Free Software, 2019. URL https://lipu.dgold.eu/original-sin. 3.0.1

[57] James Halliday. Open Source is Not Enough, May 2018. URL https://notesfrombelow.org/article/open-source-is-not-enough. 3.0.1

[58] Rob Hunter. Reclaiming the Computing Commons. *Jacobin*, May 2016. URL https://jacobin.com/2016/02/free-software-movement-richard-stallman-linux-open-source-enclosure/. 3.0.1

[59] Melody Horn. Post-Open Source, August 2020. URL https://www.boringcactus.com/2020/08/13/post-open-source.html. 3.0.1

[60] Denis Pushkarev. So, what's next?, February 2023. URL https://github.com/zloirock/core-js/blob/76f8648790efe74634874012701f387884d2c549/docs/2023-02-14-so-whats-next.md. 3.0.1

[61] Steve Marquess. Speeds and Feeds › Of Money, Responsibility, and Pride, April 2014. URL https://veridicalsystems.com/blog/of-money-responsibility-and-pride/index.html. 3.0.1

[62] Sean Gallagher. Rage-quit: Coder unpublished 17 lines of JavaScript and "broke the Internet", March 2016. URL https://arstechnica.com/information-technology/2016/03/rage-quit-coder-unpublished-17-lines-of-javascript-and-broke-the-internet/. 3.0.1

[63] Christofer Dutz. Your free trial version of "open-source" has expired, please update to a commercial plan, January 2022. URL https://github.com/chrisdutz/blog/blob/835dbf45eaa49aa153604e7e0064b29435f43554/plc4x/free-trial-expired.adoc. 3.0.1

[64] Matthew Butterick. GitHub Copilot investigation · Joseph Saveri Law Firm & Matthew Butterick, October 2022. URL https://githubcopilotinvestigation.com/. 3.0.1

[65] Matthew Butterick. GitHub Copilot litigation, November 2022. URL https://githubcopilotlitigation.com/. 3.0.1

[66] Rob O'Leary. VS Code - What's the deal with the telemetry?, April 2022. URL https://www.roboleary.net/tools/2022/04/20/vscode-telemetry.html. 3.0.1

[67] VSCodium - Open Source Binaries of VSCode. URL https://vscodium.com/. 3.0.1

[68] Philip Mirowski. The future(s) of open science. *Social Studies of Science*, 48(2):171–203, April 2018. ISSN 0306-3127. https://doi.org/10.1177/0306312718772086. URL https://doi.org/10.1177/0306312718772086. 9

[69] Doris Allhutter. Of "Working Ontologists" and "High-Quality Human Components": The Politics of Semantic Infrastructures. In *Of "Working Ontologists" and "High-Quality Human Components": The Politics of Semantic Infrastructures*, pages 326–348. Princeton University Press, May 2019. ISBN 978-0-691-19060-0. https://doi.org/10.1515/9780691190600-023. URL https://www.degruyter.com/document/doi/10.1515/9780691190600-023/html. 9, 4.1

[70] User talk:Jimbo Wales/Archive 192. *Wikipedia*, August 2015. URL https://en.wikipedia.org/w/index.php?title=User_talk:Jimbo_Wales/Archive_192&oldid=1143087052#WP_traffic_from_Google_declining. 3.0.1

[71] Roy Hinkis. Google steals 550+ million Wikipedia clicks in 6 months, traffic drop confirmed by Wiki's Jimmy Wales, August 2015. URL https://web.archive.org/web/20160114172612/http://www.similarweb.com/blog/google-steals-over-550-million-clicks-from-wikipedia-in-6-months. 3.0.1

[72] Molly White. Wikimedia timeline of events, 2014–2016, February 2016. URL http://mollywhite.net/wikimedia-timeline/. 3.0.1

[73] William Buetler. Search and Destroy: The Knowledge Engine and the Undoing of Lila Tretikov, February 2016. URL https://thewikipedian.net/2016/02/19/knowledge-engine-lila-tretikov/. 3.0.1

[74] Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1419–1428, Montréal Québec Canada, April 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. https://doi.org/10.1145/2872427.2874809. URL https://dl.acm.org/doi/10.1145/2872427.2874809. 3.0.1

[75] Wikimedia Meta-Wiki. Google - Meta. URL https://meta.wikimedia.org/wiki/Google. 3.0.1

[76] Google's stake in Wikidata and Wikipedia - Wikidata - lists.wikimedia.org, 2019. URL https://lists.wikimedia.org/hyperkitty/list/wikidata@lists.wikimedia.org/thread/KOCJRSDG57VYWQ4F2BPF7TH7R7YXGF7G/#KOCJRSDG57VYWQ4F2BPF7TH7R7YXGF7G. 3.0.1

[77] Denny Vrandečić and Markus Krötzsch. Wikidata: A Free Collaborative Knowledge Base. *Communications of the ACM*, 57(10):78–85, 2014. 3.0.1

[78] CrossRef. The Formation of CrossRef: A Short History, 2009. URL https://www.crossref.org/pdfs/CrossRef10Years.pdf. 11

[79] NISO. RP-8-2008, Journal Article Versions (JAV): Recommendations, 2008. 3.0.1

[80] NISO. RP-22-2021: Access & License Indicators, 2021. URL https://www.niso.org/publications/rp-22-2021-ali. 3.0.1

[81] Todd A. Carpenter. New Article Sharing Framework released, May 2021. URL https://scholarlykitchen.sspnet.org/2021/05/17/stm-article-sharing-framework/. 3.0.1

[82] SemTech 2011 BOF on structured data in HTML, June 2011. URL https://www.w3.org/2011/06/semtech-bof-notes.html. 3.0.1

[83] Andrew Iliadis, Amelia Acker, Wesley Stevens, and Sezgi Başak Kavakli. One schema to rule them all: How Schema.org models the world of search. *Journal of the Association for Information Science and Technology*, n/a(n/a), January 2023. ISSN 2330-1643. https://doi.org/10.1002/asi.24744. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24744. 3.0.1

[84] R.V. Guha, Dan Brickley, and Steve Macbeth. Schema.org: Evolution of Structured Data on the Web, November 2015. URL https://queue.acm.org/detail.cfm?id=2857276. 3.0.1

[85] Sandro Hawke. Notes from Q&A session at SemTech, RichSnippets session, June 2011. URL https://lists.w3.org/Archives/Public/public-vocabs/2011Jun/0001.html. 3.0.1

[86] Paul Wiegmann, Henk de Vries, and Knut Blind. Multi-Mode Standardisation: A Critical Review and a Research Agenda. *Research Policy*, 46(9):1370–1386, July 2017. ISSN 00487333. https://doi.org/10.1016/j.respol.2017.06.002. URL https://repub.eur.nl/pub/107374/. 3.0.1

[87] Marcel Heires. The International Organization for Standardization (ISO). *New Political Economy*, 13(3): 357–367, September 2008. ISSN 1356-3467. https://doi.org/10.1080/13563460802302693. URL https://doi.org/10.1080/13563460802302693. 3.0.1

[88] Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, and Honghan Wu, editors. *Exploiting Linked Data and Knowledge Graphs in Large Organisations*. Springer International Publishing, Cham, 2017. ISBN 978-3-319-45652-2 978-3-319-45654-6. https://doi.org/10.1007/978-3-319-45654-6. URL http://link.springer.com/10.1007/978-3-319-45654-6. 3.0.1, 4.1

[89] NIH Strategic Plan for Data Science. Technical report, National Institutes of Health, June 2018. URL https://web.archive.org/web/20210907014444/https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf. 3.0.2

[90] The Biomedical Data Translator Consortium. Toward A Universal Biomedical Data Translator. *Clinical and Translational Science*, 12(2):86–90, 2019. ISSN 1752-8062. https://doi.org/10.1111/cts.12591. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.12591. 3.0.2, 15, 28

[91] Christopher P. Austin, Christine M. Colvis, and Noel T. Southall. Deconstructing the Translational Tower of Babel. *Clinical and Translational Science*, 12(2):85–85, 2019. ISSN 1752-8062. https://doi.org/10.1111/cts.12595. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.12595. 16

[92] Deepak R. Unni, Sierra A. T. Moxon, Michael Bada, Matthew Brush, Richard Bruskiewich, J. Harry Caufield, Paul A. Clemons, Vlado Dancik, Michel Dumontier, Karamarie Fecho, Gustavo Glusman, Jennifer J. Hadlock, Nomi L. Harris, Arpita Joshi, Tim Putman, Guangrong Qin, Stephen A. Ramsey, Kent A. Shefchek, Harold Solbrig, Karthik Soman, Anne E. Thessen, Melissa A. Haendel, Chris Bizon, Christopher J. Mungall, and The Biomedical Data Translator Consortium. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. *Clinical and Translational Science*, n/a(n/a), May 2022. ISSN 1752-8062. https://doi.org/10.1111/cts.13302. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.13302. 3.0.2, 17

[93] Richard Bruskiewich, Deepak, Sierra Moxon, Chris Mungall, Harold Solbrig, cbizon, Matthew Brush, Kent Shefchek, Lance Hannestad, YaphetKG, Nomi Harris, bbopjenkins, diatomsRcool, Patrick Wang, Jim Balhoff, Kevin Schaper, JIWEN XIN, Phil Owen, Gregory Stupp, JervenBolleman, The Gitter Badger, Vincent Emonet, and vdancik. Biolink/biolink-model: 2.2.5. Zenodo, September 2021. URL https://zenodo.org/record/5520104. 3.0.2

[94] Nicole A. Vasilevsky, Nicolas A. Matentzoglu, Sabrina Toro, Joe E. Flack, Harshad Hegde, Deepak R. Unni, Gioconda Alyea, Joanna S. Amberger, Larry Babb, James P. Balhoff, Taylor I. Bingaman, Gully A. Burns, Tiffany J. Callahan, Leigh C. Carmody, Lauren E. Chan, George S. Chang, Michel Dumontier,

Laura E. Failla, May J. Flowers, H. A. Garrett, Dylan Gration, Tudor Groza, Marc Hanauer, Nomi L. Harris, Ingo Helbig, Jason A. Hilton, Daniel S. Himmelstein, Charles T. Hoyt, Megan S. Kane, Sebastian Köhler, David Lagorce, Martin Larralde, Antonia Lock, Irene López Santiago, Donna R. Maglott, Adriana J. Malheiro, Birgit HM Meldal, Julie A. McMurry, Moni Munoz-Torres, Tristan H. Nelson, David Ochoa, Tudor I. Oprea, David Osumi-Sutherland, Helen Parkinson, Zoë M. Pendlington, Ana Rath, Heidi L. Rehm, Lyubov Remennik, Erin R. Riggs, Paola Roncaglia, Justyne E. Ross, Marion F. Shadbolt, Kent A. Shefchek, Morgan N. Similuk, Nicholas Sioutos, Rachel Sparks, Ray Stefancsik, Ralf Stephan, Doron Stupp, Jagadish Chandrabose Sundaramurthi, Imke Tammen, Courtney L. Thaxton, Eloise Valasek, Alex H. Wagner, Danielle Welter, Patricia L. Whetzel, Lori L. Whiteman, Valerie Wood, Colleen H. Xu, Andreas Zankl, Xingmin A. Zhang, Christopher G. Chute, Peter N. Robinson, Christopher J. Mungall, Ada Hamosh, and Melissa A. Haendel. Mondo: Unifying diseases for the world, by the world, April 2022. URL https://www.medrxiv.org/content/10.1101/2022.04.13.22273750v1. 18

[95] Mungall Chris. Introduction to the BioLink datamodel, May 2018. URL https://www.slideshare.net/cmungall/introduction-to-the-biolink-datamodel. 21

[96] Karamarie Fecho, Anne E. Thessen, Sergio E. Baranzini, Chris Bizon, Jennifer J. Hadlock, Sui Huang, Ryan T. Roper, Noel Southall, Casey Ta, Paul B. Watkins, Mark D. Williams, Hao Xu, William Byrd, Vlado Dančík, Marc P. Duby, Michel Dumontier, Gustavo Glusman, Nomi L. Harris, Eugene W. Hinderer, Greg Hyde, Adam Johs, Andrew I. Su, Guangrong Qin, Qian Zhu, and The Biomedical Data Translator Consortium. Progress toward a universal biomedical data translator. *Clinical and Translational Science*, n/a (n/a), May 2022. ISSN 1752-8062. https://doi.org/10.1111/cts.13301. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cts.13301. 3.0.2

[97] John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, Adil Harroud, Lauren Sanders, Sylvain V Costes, Krish Bharat, Arjun Chakraborty, Alexander R Pico, Taline Mardirossian, Michael Keiser, Alice Tang, Josef Hardi, Yongmei Shi, Mark Musen, Sharat Israni, Sui Huang, Peter W Rose, Charlotte A Nelson, and Sergio E Baranzini. The scalable precision medicine open knowledge engine (SPOKE): A massive knowledge graph of biomedical information. *Bioinformatics*, 39(2):btad080, February 2023. ISSN 1367-4811. https://doi.org/10.1093/bioinformatics/btad080. URL https://doi.org/10.1093/bioinformatics/btad080. 3.0.2, 4

[98] Renaissance Computing Institute (RENCI). Biomedical Data Translator Platform moves to the next phase, March 2022. URL https://renci.org/blog/biomedical-data-translator-platform-moves-to-the-next-phase/. 3.0.2

[99] Renaissance Computing Institute (RENCI). Use cases show Translator's potential to expedite clinical research, March 2022. URL https://renci.org/blog/use-cases-show-translators-potential-to-expedite-clinical-research/. 3.0.2

[100] Prateek Goel, Adam J Johs, Manil Shrestha, and Rosina O Weber. Explanation Container in Case-Based Biomedical Question-Answering. page 10, September 2021. URL https://web.archive.org/web/*/https://gaia.fdi.ucm.es/events/xcbr/papers/ICCBR_2021_paper_100.pdf. 3.0.2

[101] Ruth Hailu. NIH-funded project aims to build a 'Google' for biomedical data, July 2019. URL https://www.statnews.com/2019/07/31/nih-funded-project-aims-to-build-a-google-for-biomedical-data/. 3.0.2

[102] Charlotte A Nelson, Riley Bove, Atul J Butte, and Sergio E Baranzini. Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis. *Journal of the American Medical Informatics Association : JAMIA*, 29(3):424–434, December 2021. ISSN 1067-5027. https://doi.org/10.1093/jamia/ocab270. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8800523/. 3.0.2, 22, 27

[103] Translator Consortium. Clinical Data Services Provider, April 2020. URL https://github.com/NCATSTranslator/Translator-All/blob/1c44f9a2515d239730a070201ccc7d1083c27fed/presentations/Translator_2020_Kick-Off_Presentation-Clinical_Data_Services.pdf. 3.0.2

[104] Charlotte A. Nelson, Atul J. Butte, and Sergio E. Baranzini. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. *Nature Communications*, 10(1):3045, July 2019. ISSN 2041-1723. https://doi.org/10.1038/s41467-019-11069-0. URL https://www.nature.com/articles/s41467-019-11069-0. 3.0.2

[105] University of California San Francisco. BRIDGE. URL https://bridge.ucsf.edu/. 3.0.2

[106] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The Values Encoded in Machine Learning Research, June 2022. URL http://arxiv.org/abs/2106.15590. 3.0.2

[107] Abeba Birhane. Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2), February 2021. ISSN 2666-3899. https://doi.org/10.1016/j.patter.2021.100205. URL https://www.cell.com/patterns/abstract/S2666-3899(21)00015-5. 3.0.2

[108] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: Misogyny, pornography, and malignant stereotypes, October 2021. URL http://arxiv.org/abs/2110.01963. 3.0.2

[109] A Ram, Clair A Kronk, Jacob R Eleazer, Joseph L Goulet, Cynthia A Brandt, and Karen H Wang. Transphobia, encoded: An examination of trans-specific terminology in SNOMED CT and ICD-10-CM. *Journal of the American Medical Informatics Association*, (ocab200), September 2021. ISSN 1527-974X. https://doi.org/10.1093/jamia/ocab200. URL https://doi.org/10.1093/jamia/ocab200. 3.0.2

[110] Madeline B. Deutsch. Overview of feminizing hormone therapy, June 2016. URL https://transcare.ucsf.edu/guidelines/feminizing-hormone-therapy. 3.0.2

[111] Madeline B. Deutsch. Overview of masculinizing hormone therapy, June 2016. URL https://transcare.ucsf.edu/guidelines/masculinizing-therapy. 3.0.2

[112] Finn Womack, Jason McClelland, and David Koslicki. Leveraging Distributed Biomedical Knowledge Sources to Discover Novel Uses for Known Drugs, September 2019. URL https://www.biorxiv.org/content/10.1101/765305v2. 3.0.2

[113] Chris Bizon, Steven Cox, James Balhoff, Yaphet Kebede, Patrick Wang, Kenneth Morton, Karamarie Fecho, and Alexander Tropsha. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *Journal of Chemical Information and Modeling*, 59(12):4968–4973, December 2019. ISSN 1549-9596. https://doi.org/10.1021/acs.jcim.9b00683. URL https://doi.org/10.1021/acs.jcim.9b00683. 3.0.2

[114] Florence Ashley. The Misuse of Gender Dysphoria: Toward Greater Conceptual Clarity in Transgender Health. *Perspectives on Psychological Science*, page 1745691619872987, November 2019. ISSN 1745-6916. https://doi.org/10.1177/1745691619872987. URL https://doi.org/10.1177/1745691619872987. 3.0.2

[115] MasstrichtU-IDS. Knowledge Collaboratory. Maastricht University IDS, December 2022. URL https://github.com/MaastrichtU-IDS/knowledge-collaboratory. 3.0.2

[116] Cory Doctorow. Regulating Big Tech makes them stronger, so they need competition instead. *The Economist*, June 2019. ISSN 0013-0613. URL https://www.economist.com/open-future/2019/06/06/regulating-big-tech-makes-them-stronger-so-they-need-competition-instead. 29

[117] Alan Z. Rozenshtein. Moderating the Fediverse: Content Moderation on Distributed Social Media, November 2022. URL https://papers.ssrn.com/abstract=4213674. 29

[118] Lauren Bridges. Amazon's Ring is the largest civilian surveillance network the US has ever seen. *The Guardian*, May 2021. ISSN 0261-3077. URL https://www.theguardian.com/commentisfree/2021/may/18/amazon-ring-largest-civilian-surveillance-network-us. 3.0.2

[119] AWS announces AWS Healthcare Accelerator for startups in the public sector, June 2021. URL https://aws.amazon.com/blogs/publicsector/aws-announces-healthcare-accelerator-program-startups-public-sector/. 3.0.2

[120] Jon Fingas. Amazon officially becomes a health care provider after closing purchase of One Medical, February 2023. URL https://www.engadget.com/amazon-completes-one-medical-acquisition-163431975.html. 3.0.2

[121] Rachel Lerman. Amazon built its own health-care service for employees. Now it's selling it to other companies. *Washington Post*, March 2021. ISSN 0190-8286. URL https://www.washingtonpost.com/technology/2021/03/17/amazon-healthcare-service-care-expansion/. 3.0.2

[122] Corey Quinn. You Can't Trust Amazon When It Feels Threatened, March 2021. URL https://www.lastweekinaws.com/blog/you-cant-trust-amazon-when-it-feels-threatened/. 3.0.2

[123] Susan Krashinsky. Google broke Canada's privacy laws with targeted health ads, watchdog says. *The Globe and Mail*, January 2014. URL https://www.theglobeandmail.com/technology/tech-news/google-broke-canadas-privacy-laws-with-targeted-ads-regulator-says/article16343346/. 3.0.2

[124] Krishna Bharat, Stephen Lawrence, and Mehran Sahami. Generating user information for use in targeted advertising, June 2005. URL https://patents.google.com/patent/US20050131762A1/en. 3.0.2

[125] Marc Bourreau, Cristina Caffarra, Zhijun Chen, Chongwoo Choe, Gregory S Crawford, Tomaso Duso, Christos Genakos, Paul Heidhues, Martin Peitz, Thomas Rønde, Monika Schnitzer, Nicolas Schutz, Michelle Sovinsky, Giancarlo Spagnolo, Otto Toivanen, Tommaso Valletti, and Thibaud Vergé. Google/Fitbit will monetise health data and harm consumers. (107):13, 2020. 3.0.2

[126] Rebecca Pifer. Google plans to boost Medicaid information during redeterminations, March 2023. URL https://www.healthcaredive.com/news/google-boost-medicaid-information-redeterminations/644944/. 3.0.2

[127] Yossi Matias and Corrado. Our latest health AI research updates, March 2023. URL https://blog.google/technology/health/ai-llm-medpalm-research-thecheckup/. 3.0.2

[128] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large Language Models Encode Clinical Knowledge, December 2022. URL http://arxiv.org/abs/2212.13138. 3.0.2

[129] Apple. Empowering people to live a healthier day, June 2022. URL https://www.apple.com/newsroom/pdfs/Health-Report-September-2022.pdf. 3.0.2

[130] Apple. Healthcare. URL https://www.apple.com/healthcare/. 3.0.2

[131] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, New York, NY, USA, May 2015. Association for Computing Machinery. ISBN 978-1-4503-3473-0. https://doi.org/10.1145/2740908.2742839. URL https://doi.org/10.1145/2740908.2742839. 3.0.2, 4.1.1

[132] Ying Chen, J. D. Elenee Argentinis, and Griff Weber. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. *Clinical Therapeutics*, 38(4):688–701, April 2016. ISSN 1879-114X. https://doi.org/10.1016/j.clinthera.2015.12.001. 3.0.2

[133] Chaitan Baru, Martin Halbert, Lara Campbell, Tess DeBlanc-Knowles, Jemin George, Wo Chang, Adam Pah, Douglas Maughan, Ilya Zaslavsky, Amanda Stathopoulos, Ellie Young, Kat Albrecht, Amit Sheth, Emanuel Sallinger, Katerine Osatuke, Angela Rizk-Jackson, Eric Jahn, Kenneth Berkowitz, Bandana Kar, Erica Smith, Krzystof Janowicz, Brian Handspicker, Esther Jackson, Lauren Sanders, Li Chengkai, Florence Hudson, Lilit Yeghiazarian, Cogan Shimizu, Glenn Ricart, Raschid Louiqa, Dalia Varanka, Greg Seaton, Luis Amaral, Oktie Hassenzadeh, Silviu Cucerzan, Matt Bishop, Ora Lassila, Sharat Israni, Matthew Lange, Pascal Hitzler, Ryan McGranaghan, Michael Cafarella, Paul Wormeli, Todd Bacastow, Murat Omay, Sam Klein, Ying Ding, Nariman Ammar, and Sergio Baranzini. Open Knowledge Network Roadmap: Powering The Next Data Revolution. September 2022. URL https://nsf-gov-resources.nsf.gov/2022-09/OKN%20Roadmap%20-%20Report_v03.pdf. 3.0.3

[134] National Science Foundation. NSF Convergence Accelerator awards bring together scientists, businesses, nonprofits to benefit workers, September 2019. URL https://www.nsf.gov/news/special_reports/announcements/091019.jsp. 32

[135] National Science Foundation. NSF 22-017 - Dear Colleague Letter: Encouraging Research on Open Knowledge Networks, November 2021. URL https://www.nsf.gov/pubs/2022/nsf22017/nsf22017.jsp. 3.0.3

[136] Sui Huang. NIH 1OT2TR003450-01 - EVIDARA: Automated Evidential Support from Raw Data for relay agents in Biomedical KG Queries, January 2020. URL https://reporter.nih.gov/search/PyKrY9MwK02kM4isuX9HYg/project-details/10057190. 33

[137] Sergio Baranzini, Sharat Israni, James Brase, and Sui Huang. NSF Award Search: Award # 2033569 - A1: A Multi-Scale Open Knowledge Network for Biomedicine, September 2022. URL https://www.nsf.gov/awardsearch/showAward?AWD_ID=2033569&HistoricalAwards=false. 33

[138] Krzysztof Janowicz, Pascal Hitzler, Wenwen Li, Dean Rehberger, Mark Schildhauer, Rui Zhu, Cogan Shimizu, Colby Fisher, Ling Cai, Gengchen Mai, Joseph Zalewski, Lu Zhou, Shirly Stephen, Seila Gonzalez, Bryce Mecum, Anna Carr, Andrew Schroeder, Dave Smith, Dawn Wright, Sizhe Wang, Yuanyuan Tian, Zilong Liu, Meilin Shi, Anthony D'Onofrio, Zhining Gu, and Kitty Currier. Know, Know Where, Knowwheregraph: A Densely Connected, Cross-Domain Knowledge Graph and Geo-Enrichment Service Stack for Applications in Environmental Intelligence. *AI Magazine*, 43(1):30–39, March 2022. ISSN 2371-9621, 0738-4602. https://doi.org/10.1609/aimag.v43i1.19120. URL http://ojs.aaai.org/index.php/aimagazine/article/view/19120. 33

[139] National Security Commission on Artificial Intelligence. Final Report, 2021. URL https://www.nscai.gov/2021-final-report/. 3.0.3

[140] Big Data Interagency Working Group, Subcommittee on Networking & Information Technology Research & Development, and Committee on Science & Technology Enterprise. Open Knowledge Network: Summary of the Big Data IWG Workshop, October 4-5, 2017. Technical report, November 2018. URL https://www.nitrd.gov/pubs/Open-Knowledge-Network-Workshop-Report-2018.pdf. 3.0.3, 4.1.1

[141] Mate Bioservices Inc. SPOKE Cloud, 2021. URL https://www.matebioservices.com/spoke-cloud. 3.0.3

[142] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Picador St. Martin's Press, New York, first picador edition edition, 2019. ISBN 978-1-250-21578-9. 35, 3.0.3

[143] Sarah Brayne. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press, October 2020. URL https://doi.org/10.1093/oso/9780190684099.001.0001. 3.0.3, 37, 4.1.1

[144] Information Systems Advisory Board, County of Los Angeles. CONTRACT BETWEEN THE COUNTY OF LOS ANGELES AND cFIVE SOLUTIONS, INC. FOR CONSOLIDATED CRIMINAL HISTORY REPORTING SYSTEM MAINTENANCE, SUPPORT, AND ENHANCEMENT SERVICES, September 2021. URL https://file.lacounty.gov/SDSInter/bos/supdocs/161887.pdf. 39

[145] Chief Executive Office, County of Los Angeles. Strategic Plan Goal Three: Realize Tomorrow's Government Today, May 2022. URL https://ceo.lacounty.gov/strategic-plan-goal-three/. 3.0.3

[146] Ali Farahani and Information Systems Advisory Body, County of Los Angeles. Linking Public Safety Data to the Countywide Master Data Management System, October 2016. 3.0.3

[147] Sarah Lamdan. Defund the Police, and Defund Big Data Policing, Too, June 2020. URL https://www.jurist.org/commentary/2020/06/sarah-lamdan-data-policing/. 3.0.3

[148] Matthew Guariglia. Technology Can't Predict Crime, It Can Only Weaponize Proximity to Policing, September 2020. URL https://www.eff.org/deeplinks/2020/09/technology-cant-predict-crime-it-can-only-weaponize-proximity-policing. 3.0.3, 40

[149] McGrory Kathleen, Neil Bedi, and Douglas R. Clifford. Targeted. *Tampa Bay Times*, September 2020. URL https://projects.tampabay.com/projects/2020/investigations/police-pasco-sheriff-targeted/intelligence-led-policing. 3.0.3, 40

[150] Stop LAPD Spying Coalition. Racial Terror and White Wealth in South Central. In *Automating Banishment*. November 2021. URL https://automatingbanishment.org/section/5-racial-terror-and-white-wealth-in-south-central/#operation-laser-racial-terror. 3.0.3

[151] Stop LAPD Spying Coalition. Before the Bullet Hits the Body, May 2018. URL https://stoplapdspying.org/wp-content/uploads/2018/05/Before-the-Bullet-Hits-the-Body-May-8-2018.pdf. 3.0.3

[152] Stop LAPD Spying Coalition. Letter from 28 Professors and 48 Graduate Students of UCLA to LAPD, April 2019. URL https://stoplapdspying.medium.com/on-tuesday-april-2nd-2019-twenty-eight-professors-and-forty-graduate-students-of-university-of-8ed7 3.0.3

[153] Davide Castelvecchi. Mathematicians urge colleagues to boycott police work in wake of killings. *Nature*, June 2020. https://doi.org/10.1038/d41586-020-01874-9. URL https://www.nature.com/articles/d41586-020-01874-9. 3.0.3

[154] Sun-ha Hong. Prediction as Extraction of Discretion. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 925–934, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. https://doi.org/10.1145/3531146.3533155. URL https://doi.org/10.1145/3531146.3533155. 3.0.3

[155] Stop LAPD Spying Coalition. FUCK THE POLICE, TRUST THE PEOPLE: Surveillance Bureaucracy Expands the Stalker State, June 2020. URL https://stoplapdspying.org/wp-content/uploads/2020/06/TRUST-THE-PPL-not-the-POLICE.pdf. 3.0.3

[156] Charles Tapley Hoyt, Meghan Balk, Tiffany J. Callahan, Daniel Domingo-Fernández, Melissa A. Haendel, Harshad B. Hegde, Daniel S. Himmelstein, Klas Karis, John Kunze, Tiago Lubiana, Nicolas Matentzoglu, Julie McMurry, Sierra Moxon, Christopher J. Mungall, Adriano Rutz, Deepak R. Unni, Egon Willighagen, Donald Winston, and Benjamin M. Gyori. Unifying the identification of biomedical entities with the Bioregistry. *Scientific Data*, 9(1):714, November 2022. ISSN 2052-4463. https://doi.org/10.1038/s41597-022-01807-3. URL https://www.nature.com/articles/s41597-022-01807-3. 4

[157] Freddy Lecue. On the role of knowledge graphs in explainable AI. *Semantic Web*, 11(1):41–51, January 2020. ISSN 1570-0844. https://doi.org/10.3233/SW-190374. URL https://content.iospress.com/articles/semantic-web/sw190374. 4.1.1

[158] Krisztian Balog. *Entity-Oriented Search*, volume 39 of *The Information Retrieval Series*. Springer International Publishing, Cham, 2018. ISBN 978-3-319-93933-9 978-3-319-93935-3. https://doi.org/10.1007/978-3-319-93935-3. URL http://link.springer.com/10.1007/978-3-319-93935-3. 4.1.1

[159] Patrick Stickler. Re: Monotonicity [was: Re: On Consensus] from Patrick Stickler on 2002-09-25 (w3c-rdfcore-wg@w3.org from September 2002), September 2002. URL https://lists.w3.org/Archives/Public/w3c-rdfcore-wg/2002Sep/0276.html. 4.1.1

[160] Andrei Ciortea, Simon Mayer, Fabien Gandon, Olivier Boissier, Alessandro Ricci, and Antoine Zimmermann. A Decade in Hindsight: The Missing Bridge Between Multi-Agent Systems and the World Wide Web. 2019. 4.1.1

[161] Peter McBurney and Michael Luck. The Agents Are All Busy Doing Stuff! *IEEE Intelligent Systems*, 22(4): 6–7, July 2007. ISSN 1541-1672. https://doi.org/10.1109/MIS.2007.77. URL http://ieeexplore.ieee.org/document/4287266/. 4.1.1

[162] Introducing Microsoft 365 Copilot – your copilot for work, March 2023. URL https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/. 4.1.1

[163] Krzysztof Janowicz, Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md Kamruzzaman Sarker. Neural-symbolic integration and the Semantic Web. *Semantic Web*, 11(1):3–11, January 2020. ISSN 1570-0844. https://doi.org/10.3233/SW-190368. URL https://doi.org/10.3233/SW-190368. 4.1.1

[164] Hongyu Ren, Hanjun Dai, Bo Dai, Xinyun Chen, Michihiro Yasunaga, Haitian Sun, Dale Schuurmans, Jure Leskovec, and Denny Zhou. Lego: Latent execution-guided reasoning for multi-hop question answering on knowledge graphs. In *International Conference on Machine Learning*, pages 8959–8970. PMLR, 2021. 4.1.1

[165] Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. Empowering Language Models with Knowledge Graph Reasoning for Question Answering, November 2022. URL http://arxiv.org/abs/2211.08380. 4.1.1

[166] Krisztian Balog and Tom Kenter. Personal knowledge graphs: A research agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 217–220, 2019. https://doi.org/10.1145/3341981.3344241. 4.1.1

[167] Sarah Brayne. Big Data Surveillance: The Case of Policing. *American Sociological Review*, 82(5):977–1008, October 2017. ISSN 0003-1224. https://doi.org/10.1177/0003122417725865. URL https://doi.org/10.1177/0003122417725865. 4.1.1