How Linked Data creates data-driven cultures (in business and beyond)

data.world

The (near) future of data is linked.

Three blocks from the office, Lisa's phone buzzes with a new alert: High correlation detected.

She races back to her desk and inhales the auto-generated report. She can't believe it found something this quickly. Is it a mistake? Just this morning she uploaded her dataset, the product of three long months spent sampling and compiling data from 50,000 leukemia patients in the US.

Scrolling through the report Lisa sees the boundaries of one cancer cluster line up almost perfectly with a table in a dataset about the geographic distribution of avocado orchards. No, not a mistake. This is real. And that's not the only link.

Lisa's fingers trace two trend lines across her monitor in disbelief. One line represents the number of media references to a rare fungus ravaging avocado trees. The other line is from her own leukemia diagnosis data. Same geography. Same date range. They trend upward at exactly the same rate.

Avi sets his sights on a lucrative government contract that could infuse his autonomous trucking startup, ATNMS Mobility, with enough cash to hire three more engineers and a small sales and marketing team, two things it desperately needs. The proposal is due in 60 days and the company meets or beats every stated requirement except one. The benchmarks for "disengagements"—occasions when human test supervisors have to take back control from the computer—need major improvement.

Isabela, ATNMS Mobility's head of research, has just finished a project to connect their internal datasets to standard taxonomies so that related data from academic research and government studies can be automatically discovered and considered for use in increasingly sophisticated AI models. Before the company's Linked Data initiative, this process was done on a purely ad-hoc basis. A colleague would catch a potentially usable insight in one of the journals everyone else in the space also perused, then decide if it was promising enough to justify spending hours—sometimes days—integrating the new data with the company's datasets.

More often than not, it wasn't worth it. But now, the accelerated flow of helpful external data enhances the team's knowledge faster, making its way into production at many times the previous clip. What's more, some of the most valuable data coming in is from sources not previously on the company's radar—the relevance of the data itself is a much stronger signal than whether or not it was coming from a journal or data portal on a human-compiled list of relevant sources.

Each iteration of the autonomous trucking software yields fewer disengagements during road tests, and with only 12 days until the proposal deadline, the benchmarks meet the requirements. Avi and his ATNMS Mobility colleagues feel good about their chances, and Isabela starts planning more ways her team can embed Linked Data deeper in the company's DNA.

Serendipitous connections happen when data exists in an interlinked network.

Just as the World Wide Web connects documents, which contain information rendered in human-readable natural languages, the web of the future will connect data.

Concepts described in machine-readable datasets will link to other data via common references—just as web pages are connected by hyperlinks, the navigable references to other pages. People and machines will follow these connections as easily as we browse the web today.

PUT ANOTHER WAY:

> *The Semantic Web is about two things. It is about common formats for integration and combination of data drawn from diverse sources, where on the original Web mainly concentrated on the interchange of documents. It is also about language for recording how the data relates to real world objects. That allows a person, or a machine, to start off in one database, and then move through an unending set of databases which are connected not by wires but by being about the same thing.*
>
> —Semantic Web Activity Page[1]

1. https://www.w3.org/2001/sw

This isn't a futuristic fantasy.

That quote? It's from 2001. Linked Data, or the Semantic Web, both refer to the same basic concept: we can connect data using the same architecture that powers the web. The technology has had extensive academic R&D over the last couple decades, and is already successfully deployed within large organizations that amass huge data assets—Google's Knowledge Graph[2] and Goldman Sachs' Data Lake[3] are examples of companies harnessing the power of Linked Data within private networks.

> *The push to apply this technology to the entire web isn't exactly a fringe movement, either:*
>
> *When you connect data together, you get power in a way that doesn't happen just with the web, with documents.*
>
> —Sir Tim Berners-Lee[4]

2. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
3. https://conferences.oreilly.com/strata/big-data-conference-uk-2015/public/schedule/detail/39810
4. https://www.ted.com/talks/tim_berners_lee_on_the_next_web

This is the guy who invented the web, saying “you ain’t seen nothing yet.”

If this technology exists, is mature enough to use, and is supported by one of the most influential people in the web community, why hasn’t it caught fire? What has kept it out of the mainstream?

Those of us old enough to remember the early days of the web will remember a time when converting documents to HTML and publishing them using the HTTP protocol was not the de-facto solution that it is today. The web, too, was an esoteric, academic technology with seemingly no use beyond the Ivory Tower.

Authoring HTML was vastly more difficult than writing plain text. There were no tools making it easy. Sharing documents over HTTP required technical know-how that very few people possessed. Sure, there were competing technologies like FTP, Gopher, and Usenet. If HTML was the great unifier, it wasn’t obvious at the time.

Tim Berners-Lee invented the web at CERN as a pragmatic solution to a real problem: information sharing between a tiny collective of scientists working across the globe. The first hundred thousand or so documents on the web were put there with considerable effort, and the benefit of participating in that “tiny” early web was close to nil compared to that of today’s web.

The catalyzing moment: when the true cost of sharing research information became too much to bear, innovation was the only path to a better solution.

Data: Matthew Gray of the Massachusetts Institute of Technology[5]

As with any network effect, the value of the network increased exponentially as the size of the network increased, and the costs went down as the tools improved. The advent of web content authoring platforms (e.g., WordPress, Medium) and social media reduced the costs to, essentially, zero.

For example, for your next blog post, you could hand-write the HTML code, and you could run your own HTTP server to respond to web requests for it—but why would you? Using a typical CMS, you only need to focus on the content. Anyone who can type words and click “Publish” can add information to the web.

The scope and influence of the web reached far beyond the imaginations of its innovators to become what it is today.

5. http://www.mit.edu/people/mkgray/growth/

Linked Data is at a similar inflection point.

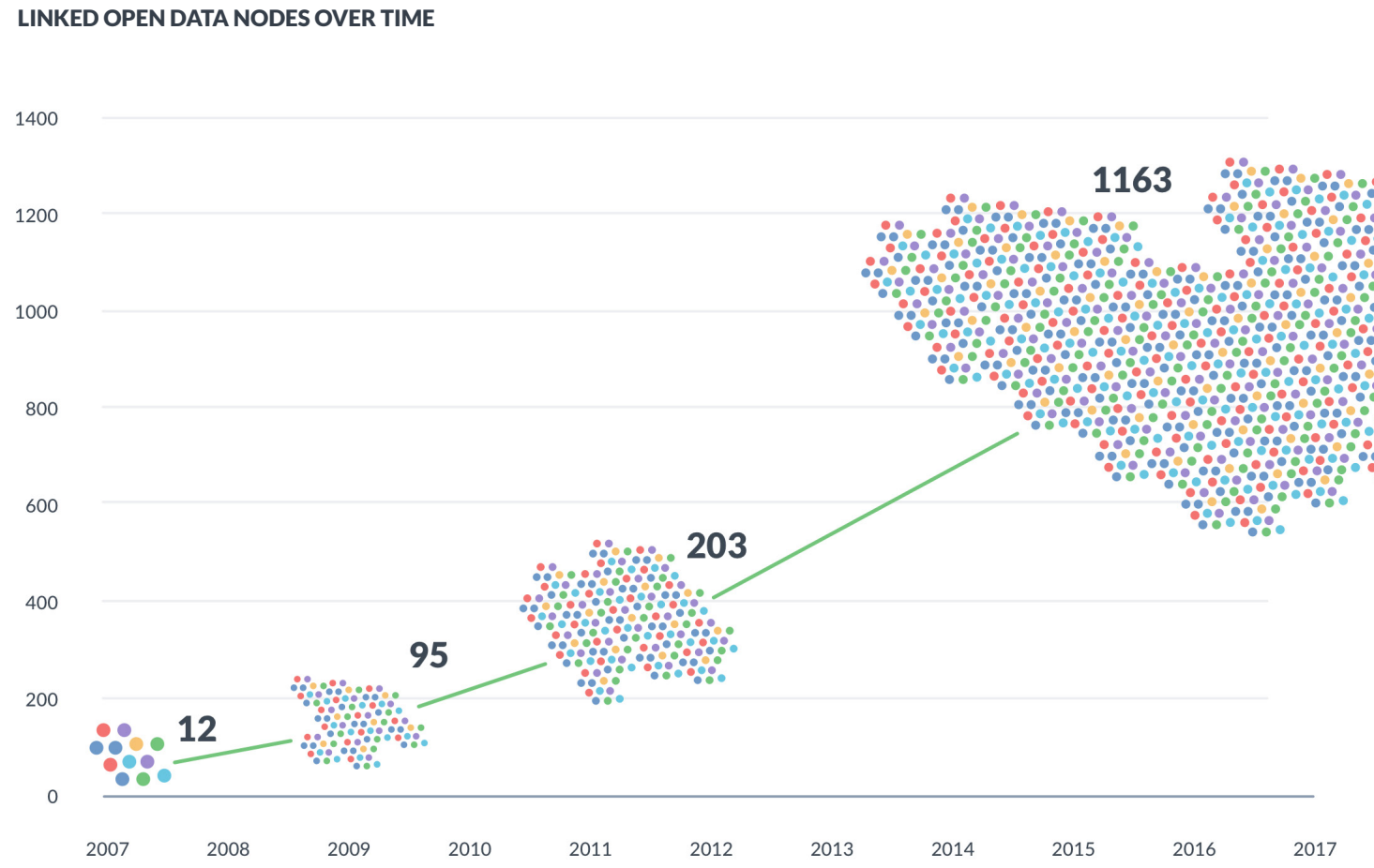
Many of the same dynamics are at play:

1. The volume and diversity of data being created is leading to a crisis, a catalyzing event: the effort required to make data interoperable is consuming way too much of the energy spent in data work. This is similar to the crisis of information sharing felt by Tim Berners-Lee and his colleagues at CERN that led to the adoption of the early web.

2. Increasingly, organizations and individuals are looking for solutions to make sense of the data explosion, and Linked Data is perfectly designed to power those solutions.

3. The web of Linked Data will continue to grow exponentially, not linearly—so each of these new steps will only accelerate the movement further.

4. A new set of tools will emerge to make the publication of Linked Data something that can be accomplished without having to become an expert in the underlying mechanisms of the Semantic Web. This is the "flywheel" effect, where the network of Linked Data will simultaneously get more valuable and cheaper to leverage.



Data: LOD-CLOUD.NET

The web of Linked Data is the natural evolution of the web, growing exponentially and adapting to the new models of information processing that have arisen as we've become more connected. The web has put previously unimaginable information resources in the palm of our hand and generates staggering amounts of data, pushing us to develop new ways to use this abundance to understand our world. Modern data science practices, especially artificial intelligence and machine learning, also drive us to reshape the way we share information on the web. Linked Data will power a lot of that. Using data will be like browsing the web—because it will be browsing the web.

There's a place for everyone in this evolution. Some will prefer to understand the underlying technologies (RDF, SPARQL, and OWL are all worth learning[6] if you want to get ahead of it). Many more will experience Linked Data through the familiar language and patterns used by mainstream data workers. More still will benefit from Linked Data without knowing anything about it, just like you're not thinking about the array of technologies that make it possible to read this very sentence.

The dream of Linked Data will crystallize while the technology fades into the background, just as it should. And once again, a decades-old idea will quietly change the world.

RDF

Resource Description Framework, a standard model for data interchange on the web.

SPARQL

SPARQL Protocol and RDF Query Language, a semantic language for databases used to retrieve and manipulate RDF data.

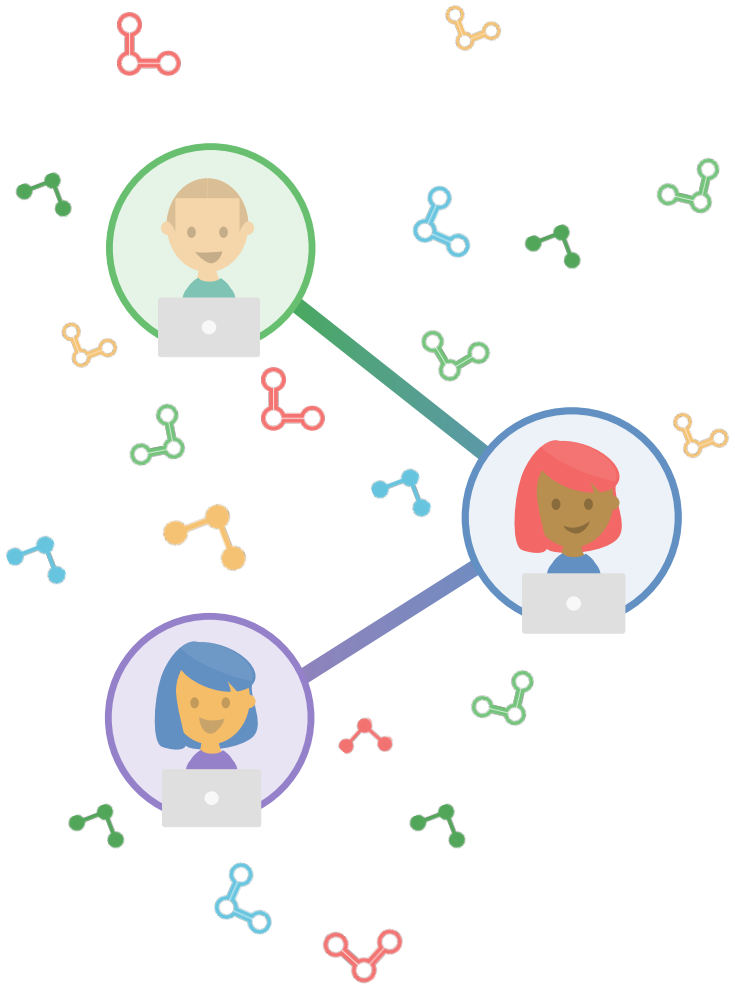
OWL

Web Ontology Language, a Semantic Web language designed to represent rich and complex knowledge about things, groups of things, and relations between things.

6. https://docs.data.world/tutorials/sparql/index.html

data.world is built on the idea that the best way to connect data is to connect people who work with data. Learn how Linked Data powers data.world.

Creating data-driven business cultures with Linked Data



Linked Data doesn't make the headlines like blockchain. Startups aren't rewriting elevator pitches to bask in its glow like with AI. But don't mistake the silence for stillness: Linked Data is ready to change business.

The first wave of companies to adopt it were established players in "semantic stronghold" industries built on massive datasets and deep R&D coffers—primarily pharma and bio sciences, aerospace, finance, and Big Tech. Now we're entering a phase during which practically any company can seize on the opportunity created by vanishing barriers to entry, a first-mover advantage worthy of Econ textbooks, compounding Big Data network effects, and a dawning realization that the energy devoted to the costliest, slowest phase of data work—preparation—can finally be reallocated to vastly more productive activities —like analysis.

Linked Data affords businesses an array of tremendous opportunities in areas like search engine visibility, lucrative recommendation engines, and scalable data integration. Several of the technology's most-heralded benefits are particularly transformative because they each provide new ways for companies to evolve into deeply data-driven cultures. In these companies, data is elemental.

If you visit a truly linked-data-driven company, what will you experience?

You will find it curiously difficult to distinguish between "traditional" data workers (analysts, data scientists, etc.) and those in other functional areas who, at other companies, are less reliant on data. The agent of change here is the unambiguous way that Linked Data represents the world. Semantic technology expert Lee Feigenbaum summarizes[7] this idea well:

> *For people, these technologies use the same language that subject-matter experts in a domain would use to talk about their data. They provide labels and descriptions intended for people, and they're not obfuscated with irrelevant IDs, codes, or abbreviations. Often, software user interfaces can be driven directly from the human-friendly descriptions of the data in RDF Schema and OWL.*

This is not to say that learning to use Linked Data today is a breeze. It's not easy to learn SPARQL or OWL, but some of the most exciting innovations in the space are closing the gap between the intuitiveness of Linked Data's structure and the relative difficulty of its exploration. Ontoforce, which has created a wonderfully accessible semantic search platform[8], is one company contributing to this effort. data.world, for its part, natively supports datasets in the RDF format, and builds an RDF model for any data that comes in a structured format (like CSV and JSON). That makes the data queryable, and each element is assigned a URI so any two datasets can be queried jointly or merged for analysis. data.world also gives people a place to collaborate on data projects using familiar file formats and their preferred tool chains, and it helps them add context to their data while it captures knowledge about its meaning and relation to other sources—the essential components of the Semantic Web. You can read more about Linked Data on data.world here[9].

Returning now to your imagined on-site visit, you will marvel at the volume and dizzying variety of data accruing harmoniously from disparate sources, flowing from team to team, integrating seamlessly with other data, ceaselessly producing unexpected insights, available to anyone at any time. This is possible because semantic technology uses URIs and shared vocabularies consistently across databases so new data can quickly and easily be added without requiring the costly, complicated changes companies must slog through when adding new data to relational databases, which are purpose-built and notoriously inflexible.

7. https://www.cmswire.com/cms/information-management/the-what-and-why-of-semantic-web-technologies-017160.php
8. https://www.youtube.com/watch?v=2_iFbPE8Wj4
9. https://meta.data.world/linked-data-on-data-world-23f5cd60ce63

All this will seem oddly familiar, resembling the web we already know in many respects. The data would be browseable and searchable by humans, crawlable and queryable by machines. Additionally, just like the web, Linked Data enjoys a remarkable network effect in that each dataset added to the network increases the incremental value of every dataset in the network. Network effects are not limited to the realm of open data; they can and do benefit organizations that are not yet publicly releasing data. The benefits of a well-maintained internal dataset, for example, can ripple across an organization, connecting to and enhancing other internal datasets, and being enhanced *by* ingesting public data like the US Census' American Community Survey. And as communities of data users grow, so does the likelihood that an individual member will have a remarkable impact on the value of the data to the rest of his or her community. For example, a subject matter expert posts a comment on a dataset that inspires the rest of the company to explore an entirely new use case for the data.

You will be inspired by the speed of and confidence in decisions, the accuracy of forecasting, and the rapid creation and adjustment of models and automated processes in response to real-time data. Much of this agility is fueled by machine learning models being deployed at a far faster pace than can be achieved without the aid of Linked Data.

URI

Uniform Resource Identifier, URIs refer to resources or information about those resources, giving a unique name to a piece of data.

Vocabulary

These define the concepts and relationships used to describe and represent an area of concern.

This is because the output of ML work is tightly correlated with the quality of input data. People who work in this area spend much of their days cleaning and preparing input data, whereas semantically-linked data has been "pre-understood" and embedded with knowledge. It will yield better results, faster, than searching through and preparing unstructured data. SiliconAngle analyst James Kobielus explains[10]:

> *Cognition, the machinery of rational thought, is empty without semantics. It would be counterproductive for the big data analytics industry to push deeper into cognitive computing without bringing the semantic Web into the heart of this new age.*

In the next section, we'll get practical with 10 useful best practices for kicking off a Linked Data program of your own.

10. https://www.infoworld.com/article/2610447/big-data/cognitive-computing-can-take-the-semantic-web-to-the-next-level.html

Launching your Linked Data program

W3C (The World Wide Web Consortium) has published a document detailing ten best practices for publishing data as Linked Open Data[11]. While this great resource is aimed at data practitioners working on open government data programs, the steps it outlines have broad applicability to anyone looking to leverage Linked Data, including those in the private sector. What follows is a summary of those best practices that reflects their broad applicability.

First steps

1: Prepare stakeholders

2: Select a dataset

The first two best practices are about building support and a business case for Linked Data. Stakeholders will need to be educated on the costs and benefits of Linked Data in order to justify investment. One of the most important considerations is selecting the datasets for publication. Your best candidates are unique datasets with a large number of potential connections to other data. These datasets are likeliest to provide a high return on investment when published as Linked Data.

These best practices apply equally well to both government agencies considering how to publish their data as Linked Data and to companies or organizations looking to leverage Linked Data to improve the discoverability and usability of data assets internally and in collaboration with external entities such as partners, suppliers, and customers.

Application independence

3: Model the data

One of the primary ways to realize the value of Linked Data is to use a model that adheres as closely as possible to the real-world concepts represented by the data, because these models are reusable in many contexts and for many applications.

In practice, this means involving a number of participants across an organization:

- a **Domain Subject Matter Expert** who deeply understands the concepts represented by the data
- a **DBA** or **Data Steward** who understands the data models and standards already in place
- a **Linked Data Subject Matter Expert** who can facilitate the modeling process and educate the rest of the group on Linked Data principles

11. https://www.w3.org/TR/ld-bp/

Usage considerations

#4 Specify an appropriate license

In the world of Linked Data published openly (sometimes referred to as LOD, Linked Open Data), it's important to declare the ownership of a dataset, and the license under which the data is provided for use. Selene Arrazolo, data.world's Lead Data Analyst, provides expert guidance on choosing the right license in this article[12].

> *If your data does not have any license terms, that means you retain all rights, and you do not authorize anyone to use, copy, distribute, share, combine, or make to changes or derivative works from it. The more open a license is, the higher the chance that others will use your data and recognize you for your work as a proponent of open data*

If your organization does not plan to publish its Linked Data , similar consideration needs to be given to the provenance and usage restrictions of data to guide internal usage.

This means recording, along with a dataset, who published it, how it was produced, and the usage restrictions on the data (who is allowed to view the data and for what purposes it is suitable according to policies).

Over the course of a dataset's lifecycle, these restrictions and suitable uses may change. Some data may start off in a tightly-controlled environment, and slowly expand its circle of influence to within an organization. Similarly, some data might be held very privately for a period of time when it contains strategic importance, and then released widely within an organization (or beyond) to support further analysis and research.

12. https://blog.data.world/what-license-should-i-use-for-my-data-a80d1ca6717b

Technical considerations

#5 Use good URIs for Linked Data
#6 Use standard vocabularies
#7 Convert data
#8 Provide machine access to data

This set of four technical best practices are equally suited to Linked Data programs, regardless of a program's degree of openness. They speak directly to producing quality Linked Data that yields its benefits fully:

- Using **good URIs** means choosing identifiers for concepts that won't change over time. Once published, the URI for an entity or concept should always refer to that thing.

- Using **standard vocabularies** means connecting your data to common, shared definitions for concepts and entities wherever they exist. This maximizes interoperability with other datasets that connect with those vocabularies.

- Data must be **converted to RDF**, the standard language for modeling Linked Data. There are several ways of **serializing** RDF data, and you might make different choices for several reasons, including human-readability or compactness of size. It's okay to share Linked Data in any of the formats, since they are all interoperable and represent an identical logical model of the data. data.world automatically builds an RDF model for structured data added to the platform.

- Finally, providing **machine access** to the data can be done in several ways. SPARQL endpoints support structured querying of Linked Data, providing HTTP access to the URIs directly allows the data to be "crawled," and providing file-based collections of data allows machines to search through it themselves.

Maintenance and support

#9 Announce new datasets

#10 Recognize the social contract

Linked Data is only useful if it can be discovered and if it is maintained for quality and stability.

- Announce the release of datasets on all relevant communication channels. Make sure that users have a way to stay aware of new assets.
- Publish structured, machine-readable catalogs of data, to make discovery accessible to applications.
- Collect feedback on the quality and usability of data. This also helps you understand demand signals for additional datasets.
- Recognize that, like software, datasets are living entities that must be maintained and cared for. The contract being entered into is that once someone has come to depend on a dataset, it will continue to exist and be maintained.

Want to try out new Linked Data features on data.world before general availability? Ping us! LinkedData@data.world

About data.world

data.world believes meaningful data work happens when all stakeholders can contribute. The platform helps your team collaborate better by connecting your organization's domain experts, decision makers, and data professionals within a shared, productive environment designed specifically for modern data teamwork. data.world is a Public Benefit Corporation headquartered in Austin, Texas. Visit data.world and follow @datadotworld and facebook.com/datadotworld for more information.

data.world