

Visualização em Linked Open Data

Roberto Stelling¹

¹Programa de Pós-graduação em Informática – Universidade Federal do Rio de Janeiro (UFRJ) – Campus da Ilha do Fundão
117.335.792 – Rio de Janeiro – RJ – Brasil

roberto@stelling.cc

Abstract. This paper reviews the data visualization for the LOD Cloud, available at <http://lod-cloud.net>, and proposes an alternate visualization for the Linked Open Data datasets, while presenting the considerations and stages that led us to the proposed design.

Resumo. Este artigo revisa a visualização de dados da LOD Cloud, disponível em <http://lod-cloud.net>, e propõe uma visualização alternativa para os dados sobre os datasets de Linked Open Data, enquanto apresenta as etapas de análise que levaram ao desenho da visualização proposta.

1. Introdução

O site <http://lod-cloud.net> publica informações sobre a LOD Cloud desde maio de 2007 [Abele, JMcCrae, Buitelaar, Jentzsch, Cyganiak 2017].

Observando o histórico das imagens curadas pelo site, percebemos que o modelo de visualização das primeiras versões da LOD Cloud foi adequado para a quantidade de datasets e de suas inter-relações até meados de 2009 quando a quantidade de nós e suas conexões começou a exceder a capacidade de expressão daquele modelo de visualização dos dados dos datasets da LOD Clout.

A visualização mais atual da LOD Cloud, ainda que muito importante como instrumento de pesquisa, não é mais tão adequada para investigação, análise e inferências sobre os dados da nuvem quanto as suas primeiras versões. O modelo de visualização foi revisto e atualizado desde 2009 mas ainda assim parece ser insuficiente para os dados atuais. Neste artigo visitamos o histórico de versões da LOD Cloud, revisamos questões importantes para a visualização dos seus dados, apresentamos o processo de desenho e projeto de uma nova visualização e exibimos o resultado da implementação do desenho proposto.

2. Linked Open Data – Dados Abertos Conectados

Segundo Tim Berners-Lee [Berners-Lee 2006], Dados *Abertos Conectados* (LOD, da sigla em inglês ou DAC, da sigla em português) são Dados Conectados, distribuídos sob uma licença aberta, que não impede a sua reutilização sem custo.

As seguintes recomendações, ou regras, em uma interpretação mais restrita, são suficientes para que dados sejam considerados Dados Conectados:

1. Utilizar URIs como nomes de objetos (coisas)
2. Usar URIs HTTP de forma que seja possível encontrar estes nomes
3. Prover informações úteis quando da busca por URIs, utilizando os padrões (RDF, SPARQL)

4. Incluir enlaces para outras URIs, de forma a ser possível descobrir mais informações.

Ainda de acordo com Berners-Lee [Berners-Lee 2006], é necessário seguir estas recomendações para que os dados sejam considerados Dados Conectados, mas sem licença aberta de utilização, os repositórios não podem ser considerados Dados *Abertos* Conectados.

Sob um esquema hierárquico de estrelas, proposto por Berners-Lee [Berners-Lee 2006], temos:

As 5 estrelas de Dados Abertos Conectados

- ★ Disponível na web (em qualquer formato), mas com *licença aberta*
- ★★ Disponível como dado estruturado legível por máquina (e.g. excel)
- ★★★ Como acima, mas em formato não proprietário (e.g. CSV, ao invés de excel)
- ★★★★ Utilizando padrões abertos da W3C (RDF e SPARQL)
- ★★★★★ Dados conectados a outros dados abertos

A classificação de 5 estrelas permite um claro entendimento da hierarquia natural entre as aplicações de dados abertos, desde a publicação online com licença de utilização sem custo até o momento em que podem ser classificados como Dados Abertos Conectados. Partindo da simples publicação de dados para o consumo público em formatos não estruturados, como PDF ou imagens sem anotação, até datasets serializados em RDF conectados a outros datasets.

Essa hierarquia também fornece uma trilha natural de complexidade que pode ser seguida por qualquer agente que deseje publicar e compartilhar seus dados dentro do padrão de Dados Abertos Conectados.

Dados abertos conectados são os blocos de construção da Web Semântica.

3. A LOD Cloud

Desde maio de 2007, o site The Linking Open Data cloud diagram, vem publicando e mantendo atualizada uma visualização do diagrama da LOD Cloud. A visualização mostra datasets que foram publicados no formato de Dados Conectados, por participantes da comunidade do projeto Linking Open Data e outros indivíduos e organizações. O diagrama é baseado nos metadados coletados e curados por contribuintes do Data Hub. Atualmente o diagrama é mantido por Andrejs Abele e John McCrae [Abele, McCrae, Buitelaar, Jentzsch e Cyganiak 2017].

Além do diagrama mais atual da LOD Cloud, a página mantém um registro histórico do diagrama desde a sua primeira versão, como mostrado na Figura 1: A LOD Cloud em maio de 2007.

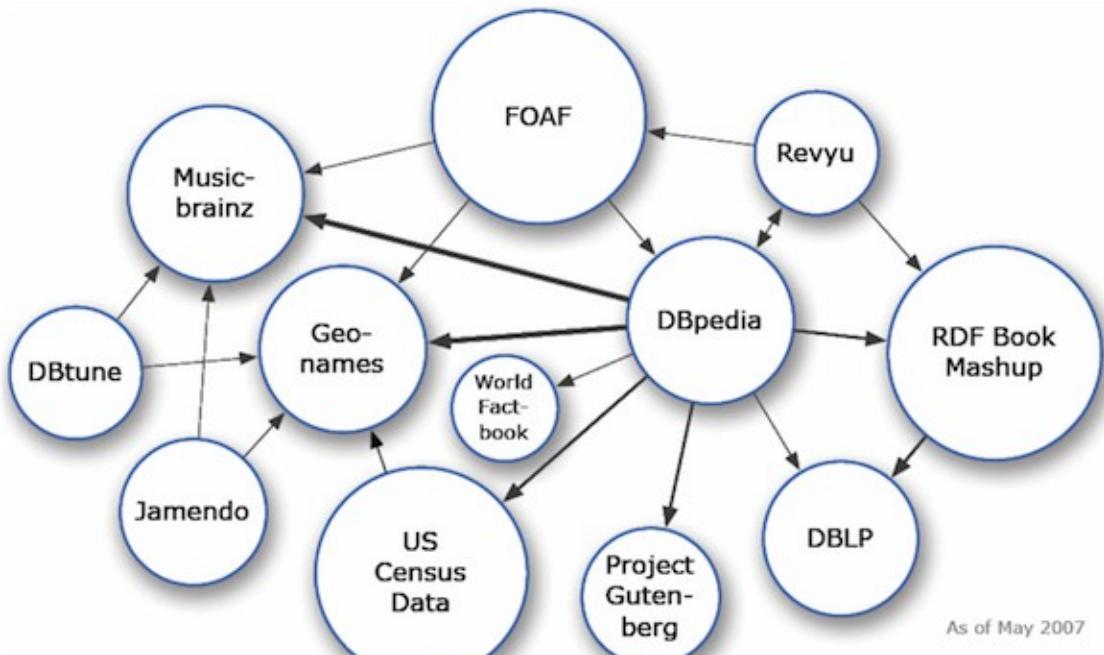


Figura 1: A LOD Cloud em maio de 2007

Fonte: The Linking Open Data cloud diagram, 2017

Com o passar do tempo e com o crescimento dos datasets representados na LOD Cloud o diagrama evoluiu até a versão atual, mostrada na Figura 2: A LOD Cloud em agosto de 2017.

Um razoável esforço foi investido não apenas em manter o diagrama mas também em representar de forma adequada os datasets que fazem parte da nuvem.

É claro que se espera que, com o passar do tempo, não seja mais possível representar, em apenas uma visão, toda a LOD Cloud. Como comparação, observamos que pelo crescimento da internet, não é possível representar toda a rede em uma só imagem, identificando cada um dos nós existentes. Por outro lado há visualizações mostrando a topologia da Internet, como a World Internet Topology, de 2007.

Alguns dos dilemas e dificuldades de representar visualmente um conjunto de dados interconectados estão explícitos na própria sequência histórica do diagrama da LOD Cloud. Por exemplo, em março de 2009 os limites de expressividade visual do modelo original foram expandidos com uma codificação por cores, que visava identificar as diversas áreas dos datasets classificados, como mostrado na Figura 3: LOD Cloud em março de 2009, versão colorida. Nessa nova representação, se percebe um crescimento marcante não apenas no número de datasets em relação a diagramas anteriores, mas também na quantidade de interconexões, quando comparamos com a versão de maio de 2007.

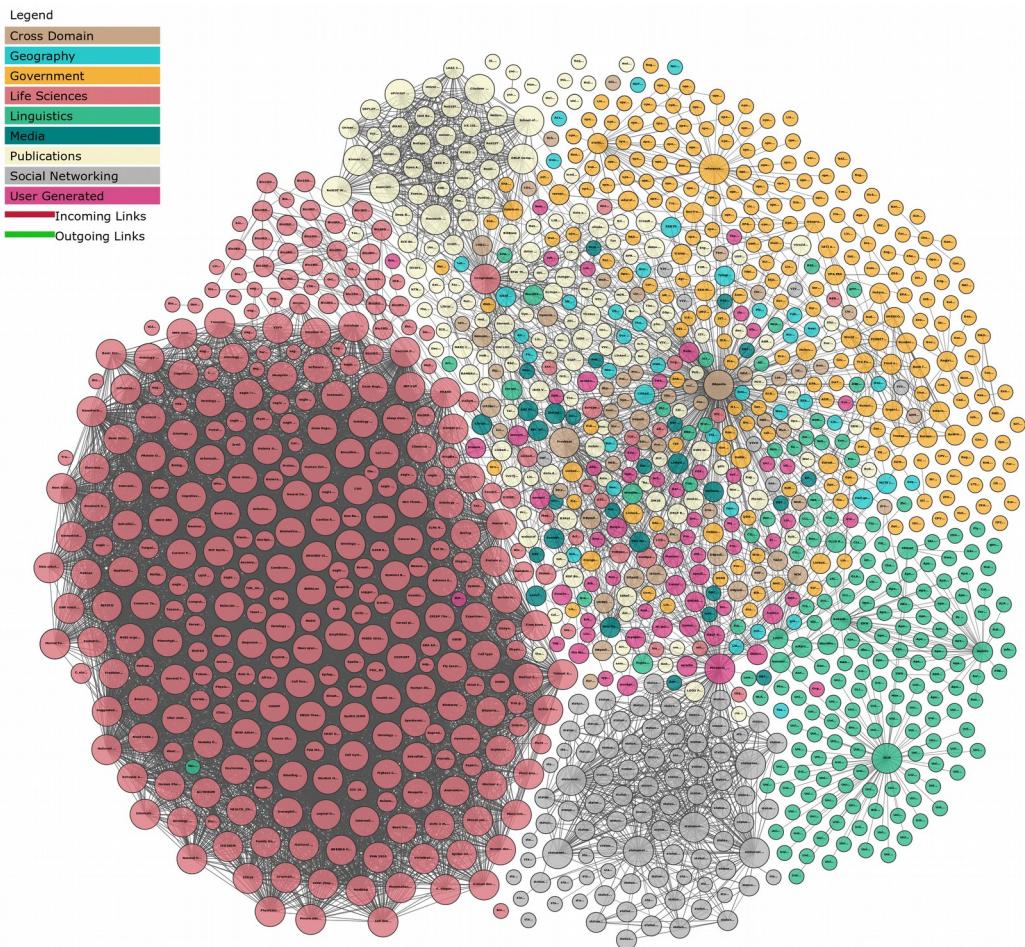


Figura 2: A LOD Cloud em agosto de 2017

Fonte: The Linking Open Data cloud diagram, 2017

Em agosto de 2014, o modelo com cores parece ter sido exigido ao extremo, como mostrado na Figura 4: LOD Cloud em agosto de 2014, onde já não é mais possível identificar as conexões de vários dos nós; os enlaces em algumas regiões formam um emaranhado com mais ruído que informação. É importante notar que a primeira versão da nuvem, de maio de 2007, possui apenas 12 datasets, enquanto a versão de agosto de 2014 possui 570. O diagrama foi redesenhado em janeiro de 2017, incluindo a distinção visual, na versão SVG, dos enlaces de entrada, em vermelho, com os enlaces de saída, em verde. Nesse momento a LOD Cloud contava com 1.146 datasets e dezenas de milhares de interconexões entre eles.

A versão mais atual, na Figura 2: A LOD Cloud em agosto de 2017, possui nada menos que 1.163 datasets, divididos em 9 categorias (Cross Domain, Geography, Government, Life Sciences, Linguistics, Media, Publications, Social Networking e User Generated) e foi publicado, pela primeira vez, com informações adicionais em JSON e CSV, além dos formatos de imagem já existentes nas versões anteriores (PNG, PDF e SVG).

Como interpretar o diagrama da Figura 2: A LOD Cloud em agosto de 2017?

Segundo os autores, a imagem mostra datasets que estão publicados no formato Linked Data (Dados Conectados) e que são interligados com outros datasets da nuvem.

O tamanho dos círculos corresponde ao número de nós conectados em cada dataset. Os tamanhos são calculados de acordo com os datasets conectados no diagrama.

Tamanho do círculo	Contagem de enlaces
Grande	>100
Médio	50-100
Pequeno	1-50

Uma linha indica a existência de ao menos um enlace entre os dois datasets. Um enlace, para os propósitos do diagrama, é uma tripla RDF onde as URIs do sujeito e objeto estão em namespaces de dataset distintos.

Na versão interativa (SVG), a cor da linha indica a direção do enlace. Por exemplo, se um enlace entre A e B é verde, isso significa que o dataset A possui triplas RDF que usam identificadores de B e, se o enlace é vermelho, isso significa que o dataset B contém triplas RD que utilizam identificadores de A.

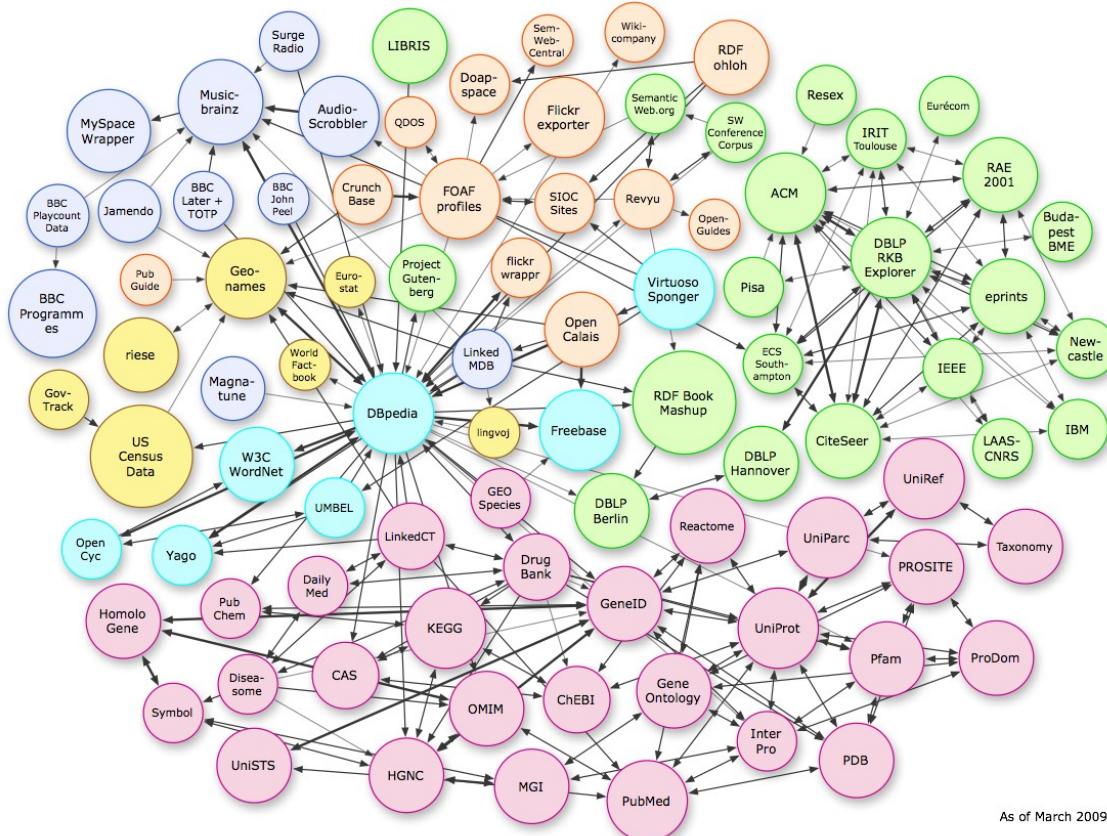


Figura 3: LOD Cloud em março de 2009, versão colorida

Fonte: The Linking Open Data cloud diagram, 2017

A inclusão de cores na direção dos enlaces da LOD Cloud e a criação da sua primeira versão interativa, em janeiro de 2017, disponível em [Interactive Cloud Image 2017-01-26], foram passos importantes no enriquecimento da representação da LOD Cloud,. Porém estes esforços já se mostram insuficientes, e talvez inadequados, para representar

os datasets e suas conexões em apenas uma imagem, dado o estado atual da LOD Cloud

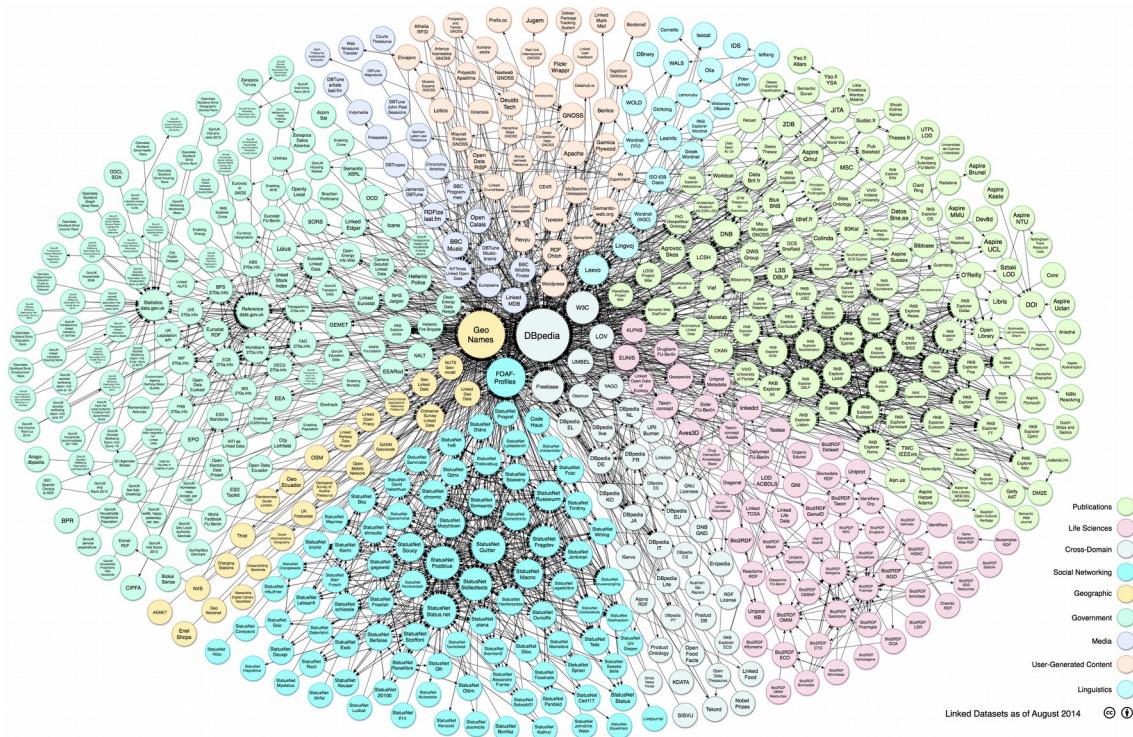


Figura 4: LOD Cloud em agosto de 2014

Fonte: The Linking Open Data cloud diagram, 2017

4. Objetivos da visualização de datasets

A partir da análise da LOD Cloud, entendemos que o diagrama da nuvem tem, pelo menos, os seguintes objetivos: identificar os datasets de dados abertos conectados, quantificar os datasets, identificar conexões entre os datasets, identificar os tipos de conexões entre os datasets e identificar os domínios dos datasets.

Que outras informações podem ser interessantes incluir na visualização ? (os próprios autores reconhecem que é possível visualizar a nuvem sob outras perspectivas [Abele, McCrae, Buitelaar, Jentzsch e Cyganiak 2017]).

Listamos algumas destas informações adicionais: taxa de erros de um dataset (índice de referências inválidas), última atualização de um dataset, taxa de atualização de um dataset (a frequência com que o dataset é atualizado), autor do dataset e outras métricas que sejam importantes.

Algumas destas informações, como autor e data de última atualização, estão disponíveis nos dados fornecidos pelo grupo responsável pela LOD Cloud, mas outros teriam que ser coletados a partir dos datasets originais e do seu histórico. A coleta de informações adicionais, ainda que interessante e útil, está fora do escopo deste trabalho.

Outras questões importantes sobre a visualização são: é possível mostrar todas as variáveis em uma só imagem/gráfico? Quais os relacionamentos que podem ser exibidos? Hierarquia, subconjuntos, relações parte → todo, entre outros.

5. LOD Target

Nesta e nas próximas seções faremos uma revisão dos objetivos de uma representação visual dos datasets de Dados Abertos Conectados, examinaremos os dados disponíveis para a construção da representação, revisaremos possibilidades e alternativas de desenho da visualização e proporemos uma alternativa para o formato atual da LOD Cloud, que chamamos de LOD Target.

5.1 Objetivos da visualização

O ponto de partida do processo de desenho de uma visualização é a definição dos objetivos da visualização.

Para o LOD Target desejamos: exibir milhares de datasets, identificar as conexões entre datasets, identificar a quantidade de referências entre datasets, identificar tamanho dos datasets e identificar as categorias a que pertencem os datasets.

5.2 Alternativas de representação visual

A LOD Cloud é uma visualização de datasets de Dados Abertos Conectados e suas referências (pela existência de triplas com URIs de namespaces de datasets distintos em um dataset). Essencialmente estamos lidando com um grafo direcionado, que é usualmente visualizado de forma gráfica com imagens similares à Figura 5: Exemplo de visualização de um grafo

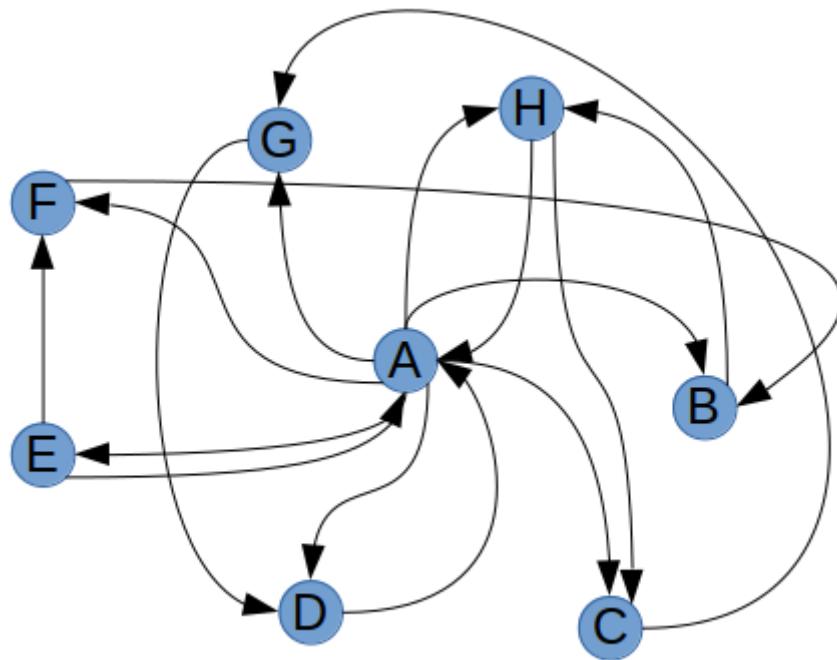


Figura 5: Exemplo de visualização de um grafo

Apesar de esta ser a representação mais usual, há outras alternativas para representar redes, em particular redes com alto índice de conexão.

Segundo Tamara Munzner [Munzner 2014], a forma mais comum de representar árvore e redes são os diagramas nós-enlace, onde nós são representados com marcadores de ponto e enlaces são representados como marcadores de linha. Porém há outras formas

de representação de grafos, cada uma com características e expressividade próprias. Vejamos algumas.

Diagramas de força (Force-Directed Placement), como mostrado na Figura 6: Exemplo de diagrama de força, são efetivos em vários casos de visualização de redes e foram certamente a base das visualizações mais recentes da LOD Cloud. De acordo com Tamara Munzner [Munzner 2014], diagramas de força tem as seguintes características:

Tabela 1: Características de diagramas de força

Formato	Diagrama de força
Objeto: Dados	Rede.
Como: Codificação	Marcadores de pontos para nós, marcadores de conexão para enlaces.
Por quê: Tarefas	Explorar topologia, localizar caminhos.
Escala	Nós: dezenas/centenas. Enlaces: centenas. Densidade nó/enlace: $E < 4N$

Adaptado de [Munzner 2014]

A tabela acima já expõe alguns problemas que ocorrem ao utilizar diagramas de força para representar os datasets de Dados Abertos Conectados: a escala do número de nós, número de enlaces e muito provavelmente a densidade de nós/enlaces (que está visível pelo emaranhado de linhas na Figura 2: A LOD Cloud em agosto de 2017).

Outra alternativa para representação de redes, quando o número de enlaces é alto, é a adoção de uma visão hierárquica de diagramas de força: diagrama de força multinível.

Tabela 2: Características de diagramas de força multinível

Formato	Diagrama de força multinível
Objeto: Dados	Rede.
Objetos: Derivação	Hierarquia de clusters sobre a rede original.
Como: Codificação	Marcadores de pontos para nós, marcadores de conexão para enlaces.
Por quê: Tarefas	Explorar topologia, localizar caminhos e clusters.
Escala	Nós: 1.000 a 10.000. Enlaces: 1.000 a 10.000. Densidade nó/enlace: $E < 4N$

Adaptado de [Munzner 2014]

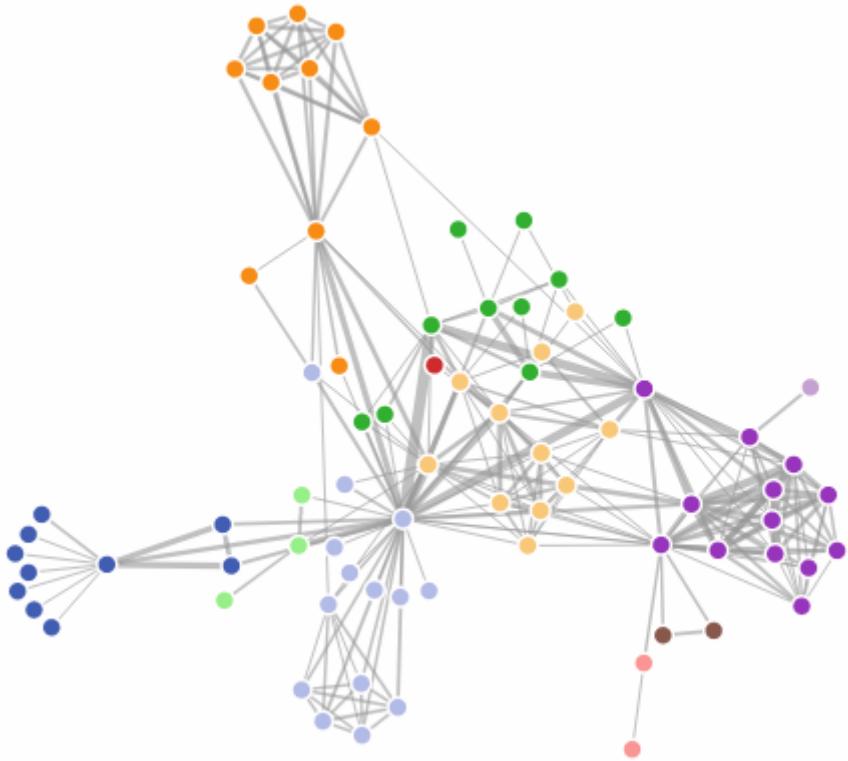


Figura 6: Exemplo de diagrama de força

Fonte: Force-Directed Graph

Embora diagramas de força multinível permitam visualizar, de forma adequada, redes com a quantidade de nós da LOD Cloud, vemos que a quantidade de enlaces e a densidade de nó/enlace da LOD Cloud são superiores à escala adequada para estes diagramas.

Uma terceira alternativa, que não sofre do problema de densidade nó/enlace são as matrizes de adjacência.

Tabela 3: Características de matrizes de adjacência

Formato	Matriz de adjacência
Objeto: Dados	Rede.
Objeto: Derivação	Tabela: nós da rede como chaves, estados dos enlaces entre nós como valores.
Como: Codificação	Marcadores de área em matrizes 2D.
Escala	Nós: 1.000. Enlaces: 1.000.000

Adaptado de [Munzner 2014]

Embora matrizes de adjacência sejam representações de redes pouco frequentes, encontramos belos exemplos interativos, como o mostrado na Figura 7: Matriz de adjacência para os personagens de Les Misérables, desenvolvida pelo autor do pacote de visualização de dados D3 para JavaScript, Mike Bostock.

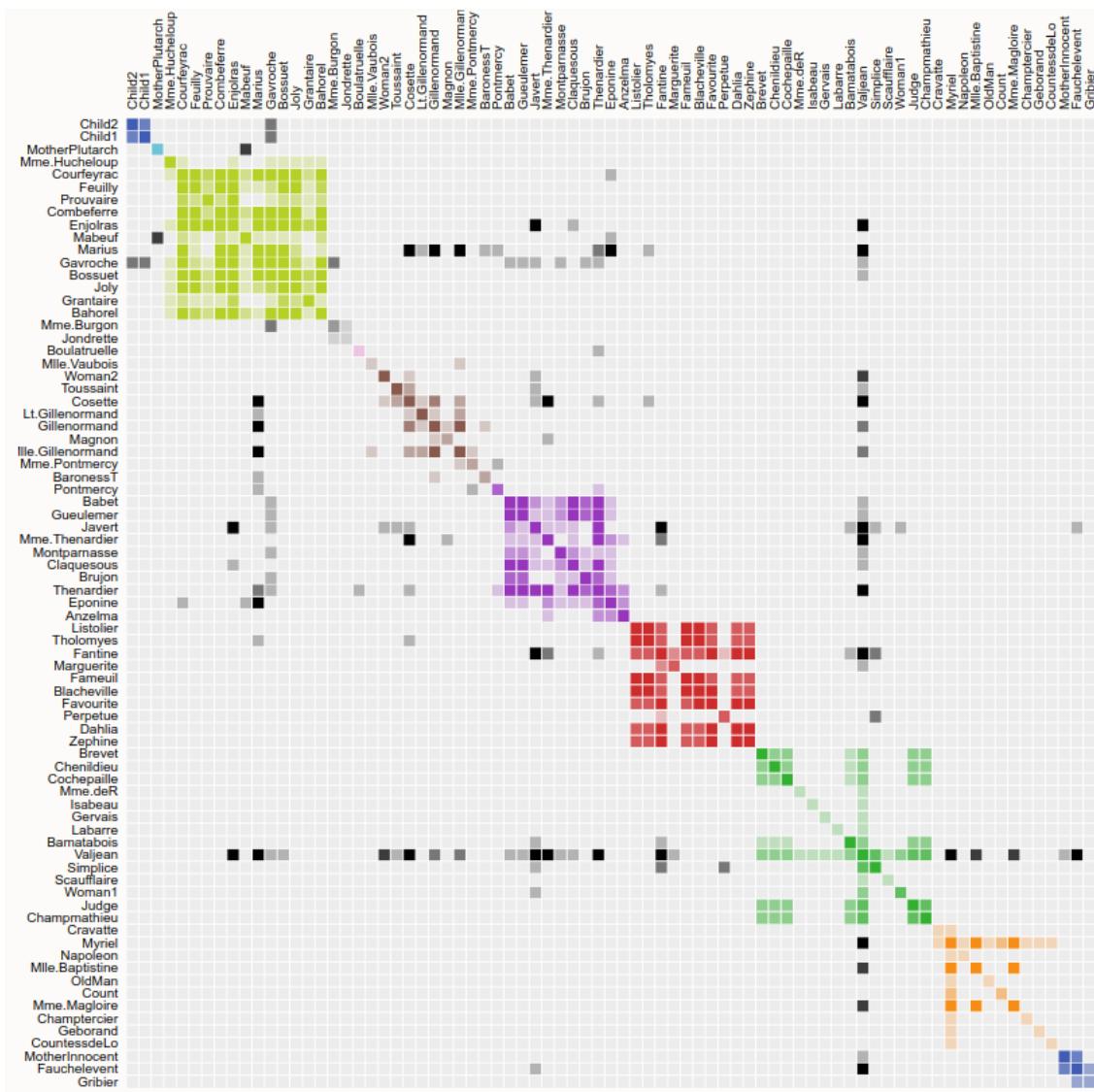


Figura 7: Matriz de adjacência para os personagens de *Les Misérables*

Fonte: *Les Misérables Co-occurrence*

Matrizes de adjacência são realmente efetivas quando há alternativas de reordenação da matriz [Munzner 2014], em especial quando é possível reordená-las por clusters, como na Figura 7: Matriz de adjacência para os personagens de *Les Misérables*. Outras ordenações das chaves da matriz, como ordenação por frequência, dão mais informações sobre os dados, porém o ponto fraco das matrizes de adjacência está na limitação da investigação de características topológicas da rede subjacente, já que os enlaces são mostrados de forma indireta pelas células da matriz.

Outras formas alternativas de representação de redes são:

Gráfico bipartido, mostrado na Figura 8: Exemplo de gráfico bipartido, onde os nós são representados como chaves e os enlaces como linhas entre as chaves dos dois lados. Cores e expressura podem ser utilizados para representar o “tamanho” dos enlaces. Conexões direcionadas são entendidas como saindo dos nós do lado esquerdo do gráfico e entrando no lado direito.

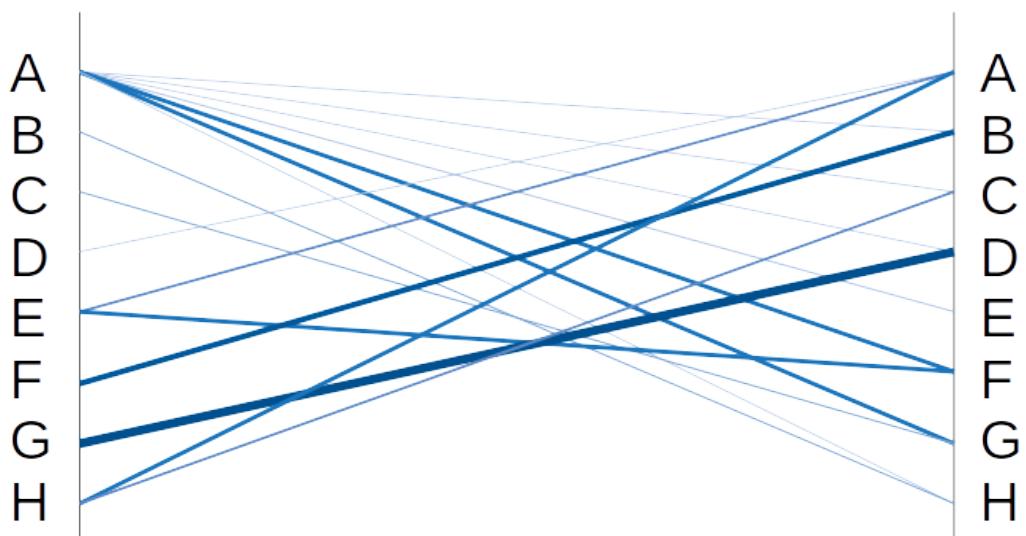


Figura 8: Exemplo de gráfico bipartido

O gráfico bipartido circular, mostrado na Figura 9: Exemplo de gráfico bipartido circular, é similar em expressividade ao anterior, porém os nós estão dispostos ao redor de um círculo e não em duas linhas paralelas.

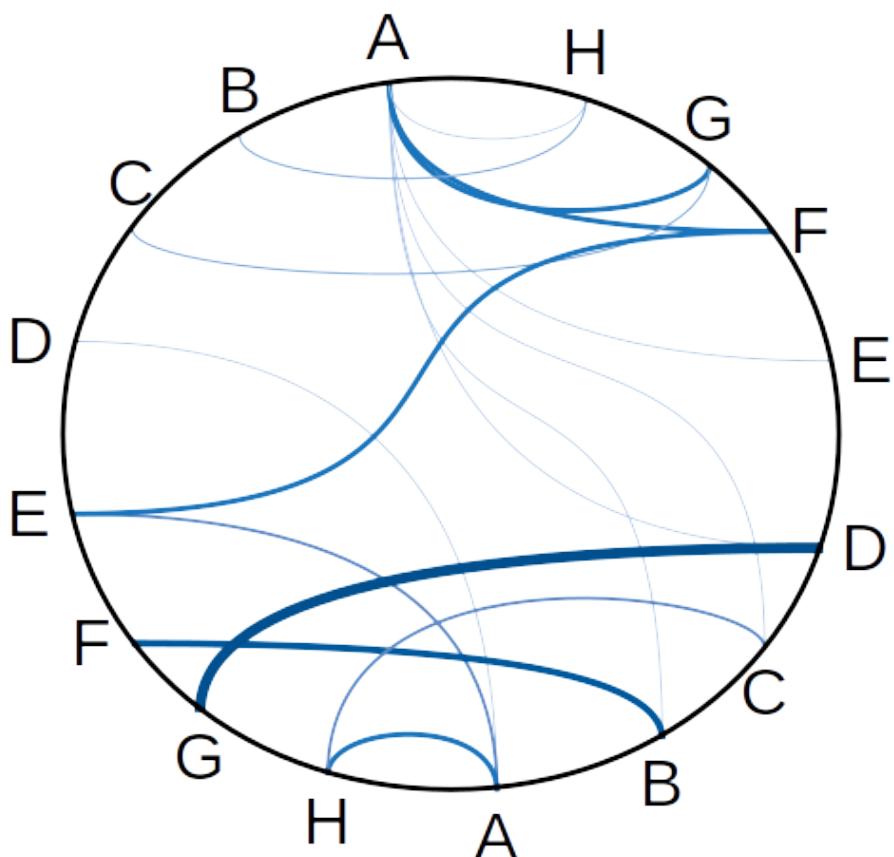


Figura 9: Exemplo de gráfico bipartido circular

Da mesma forma que o gráfico bipartido, os enlaces entre nós são entendidos como saindo dos nós à esquerda e entrando nos nós à direita. As duas figuras acima representam a mesma rede com seus nós e conexões.

5.3 Dados da LOD Cloud

Além das próprias imagens em SVG e PNG, é possível visualizar os dados da LOD Cloud em uma representação JSON da rede e também com um arquivo de apoio em TSV (valores separado por tabs).

O arquivo JSON é dividido em duas partes. A primeira descreve os nós (Figura 10: JSON - nós).

```
{ "nodes": [ { "id": 0, "size": 1, "name": "warsampo", "group": "Publications", "title": "WarSampo\n Last modified: 2017-08-21\n Triangles: 9052788", "triples": 9052788 }, { "id": 1, "size": 5, "name": "b3kat", "group": "Publications", "title": "B3Kat - Library Union Catalogues of Bavaria, Berlin and Brandenburg\n Creator: Bavarian State Library, Bavarian Library Union, Cooperative Library Net", "triples": 1008672450 }, { "id": 2, "size": 1, "name": "dbkwik", "group": "Cross domain", "title": "DBKWiK\n Creator: Alexandra Hofmann, Samresh Perchani, Jan Portisch,Sven Hertling, Heiko Paulheim\n Last modified: 2017-07-28\n Triangles: 26694082", "triples": 26694082 } ] }
```

Figura 10: JSON - nós

Enquanto a segunda descreve os enlaces (Figura 11: JSON - enlaces)

```
"links": [ { "source": 0, "weight": 900, "target": 13 }, { "source": 1, "weight": 74722718, "target": 3 }, { "source": 1, "weight": 34624, "target": 192 }, { "source": 2, "weight": 768902, "target": 13 }, { "source": 3, "weight": 80504, "target": 13 }, { "source": 3, "weight": 15723, "target": 178 } ] }
```

Figura 11: JSON - enlaces

Já o arquivo TSV possui informações adicionais sobre os nós (Figura 12: Arquivo TSV)

Title	Datahub URL	Last Updated	Domain
DBkwik	https://old.datahub.io/dataset/dbkwik	2017-07-28	Cross_domain
Gemeinsame Normdatei (GND)	https://old.datahub.io/dataset/dnb-gemeinsame-normdatei	2017-07-21	Cross_domain
Muninn World War I	https://old.datahub.io/dataset/muninn-world-war-i	2017-06-22	Cross_domain
DBpedia	https://old.datahub.io/dataset/dbpedia	2017-06-20	Cross_domain
Imagesnippets Image Descriptions	https://old.datahub.io/dataset/imagesnippets	2017-06-14	Cross_domain
WebIsALOD	https://old.datahub.io/dataset/webisalod	2017-06-14	Cross_domain
Enipedia - Energy Industry Data	https://old.datahub.io/dataset/enipedia	2016-07-30	Cross_domain
Uberlinc.org	https://old.datahub.io/dataset/uberlinc	2016-07-30	Cross_domain

Figura 12: Arquivo TSV

A partir destes arquivos realizamos um primeiro sumário dos dados da LOD Cloud:

- 1.163 nós
- 15.655 enlaces (de 1.352.579 enlaces possíveis – 1,16% de ocupação de uma matriz de adjacência → matrix altamente esparsa)
- Relação Enlaces/Nós ~13,46
- 9 grupos
- Maior grupo: Life Sciences com 332 nós
- Menor grupo: Media, com 27 nós
- Vários nós sem informação de número de enlaces

5.4 Explorando os dados visualmente

Com os dados da LOD Cloud em mãos inicia-se a fase seguinte do processo de desenho da visualização: a exploração visual dos dados.

Há várias alternativas para esta fase e vários caminhos podem ser adotados em função dos interesses e inclinações de cada projetista. No caso em particular da nuvem, achamos interessante visualizar a rede como um campo de força, sem representação dos tamanhos dos datasets, sem representação da direção dos enlaces e sem representação do tamanho das conexões (quantidade de triplas referenciadas entre os nós).

Essas simplificações permitem explorar questões de topologia e arquitetura da rede como um todo que talvez estejam mascaradas por excesso de informação no diagrama da LOD Cloud.

Pela imagem resultante, mostrada na Figura 13: LOD Cloud como campo de força, percebemos várias características dos datasets de Dados Abertos Conectados:

Há uma grande nuvem ao lado esquerdo (a nuvem é formada pelos nós e, principalmente, pelos enlaces, que de tão densos chegam a perder a capacidade informativa), neste ponto ainda não sabemos do que se trata esta parte da nuvem mas, por comparação com a Figura 2: A LOD Cloud em agosto de 2017, é razoável inferir que sejam todos, ou quase todos, da categoria Life Sciences.

Existe na imagem um ponto central, que parece referenciar e/ou ser referenciado por um grande número de datasets (muito provavelmente este nó representa a Dbpedia). Há outros cliques e clusters mas estas duas características são as mais marcantes da imagem.



Figura 13: LOD Cloud como campo de força

O passo seguinte foi analisar cada uma das subcategorias da LOD Cloud, de acordo com suas quantidades de nós e enlaces dentro da própria categoria. Os nove gráficos podem ser vistos na Tabela 4: Arquitetura das 9 categorias de Dados abertos Conectados.

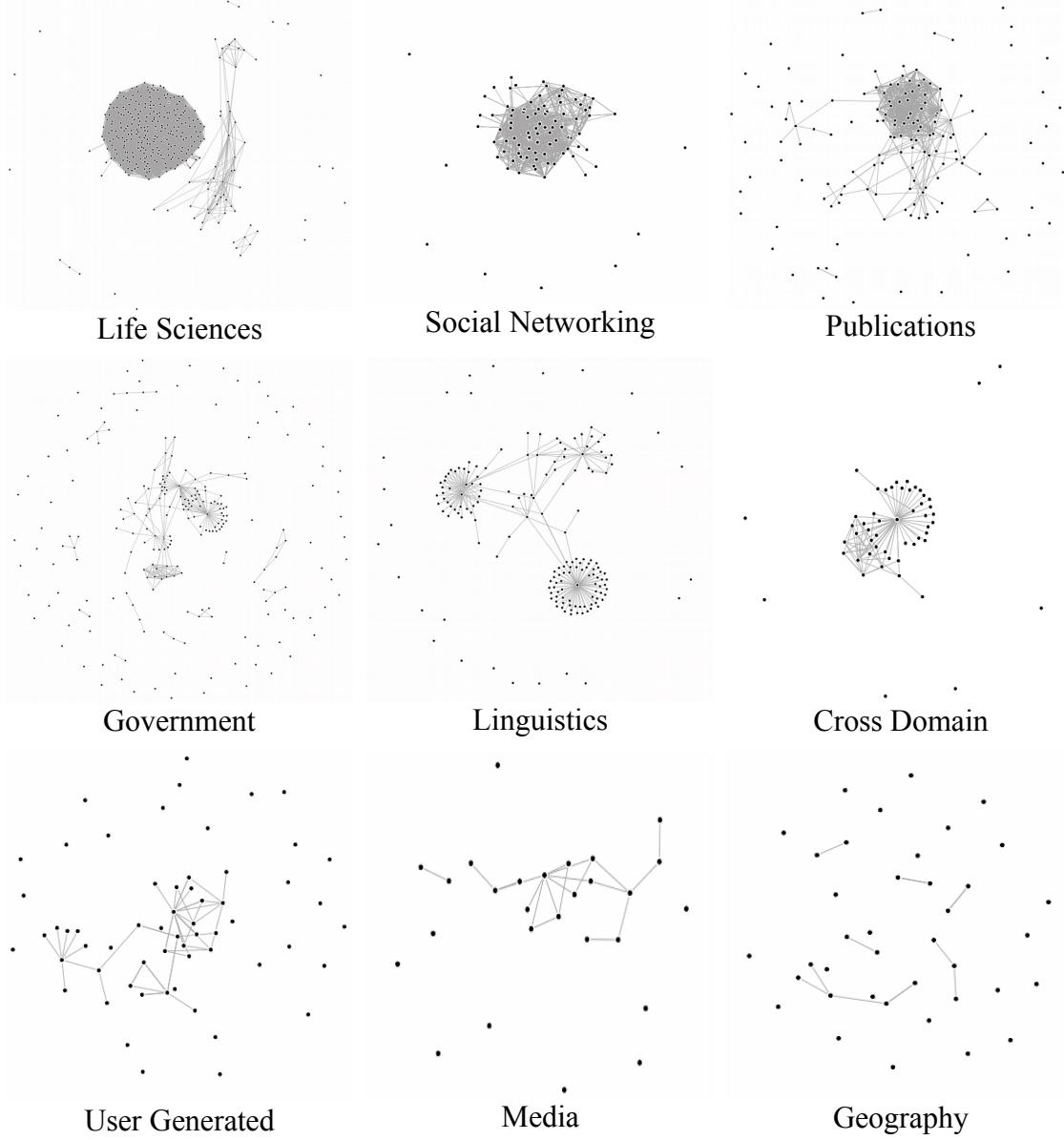
Da esquerda para direita e de cima para baixo as categorias mostradas, em ordem de densidade de enlaces/nós, são:

- Life Sciences: 332 nós, 12.638 enlaces
- Social Networking: 93 nós, 782 enlaces
- Publications: 156 nós, 643 enlaces
- Government: 229 nós, 241 enlaces
- Linguistics: 174 nós, 191 enlaces
- Cross Domain: 53 nós, 80 enlaces
- User Generated: 60 nós, 49 enlaces

- Media: 27 nós, 20 enlaces
- Geography: 39 nós, 13 enlaces

Algumas categorias, como Geography e Media, são esparsas o suficiente para serem entendidas até mesmo com os thumbnails abaixo, enquanto outras, como Life Sciences e Social Networking, são densas demais, em número de nós e enlaces, e excedem a capacidade expressiva do modelo de representação gráfica atual.

Tabela 4: Arquitetura das 9 categorias de Dados Abertos Conectados



Esta primeira visão de toda a rede e seus “pequenos múltiplos” divididos pelas categorias já indicam que uma representação gráfica que inclua os enlaces ficará tomada por uma nuvem densa de enlaces, a clássica “bola de pelos”, tão comum em representações de conexões em redes sociais.

Nesse ponto executamos um segundo experimento visual, ainda utilizando a representação de campo de força, mas agora incluindo informações de categoria, tamanho dos nós (número de triplas), enlaces e tamanho dos enlaces. Na primeira tentativa dimensionamos o tamanho dos nós proporcionais ao \log_{10} do número de triplas. O resultado, mostrado em Figura 14: DAC - Raio do nó = $\log_{10}(\#\text{triplas})$, se mostrou insatisfatório sob o ponto de vista de exploração e compactação da informação.

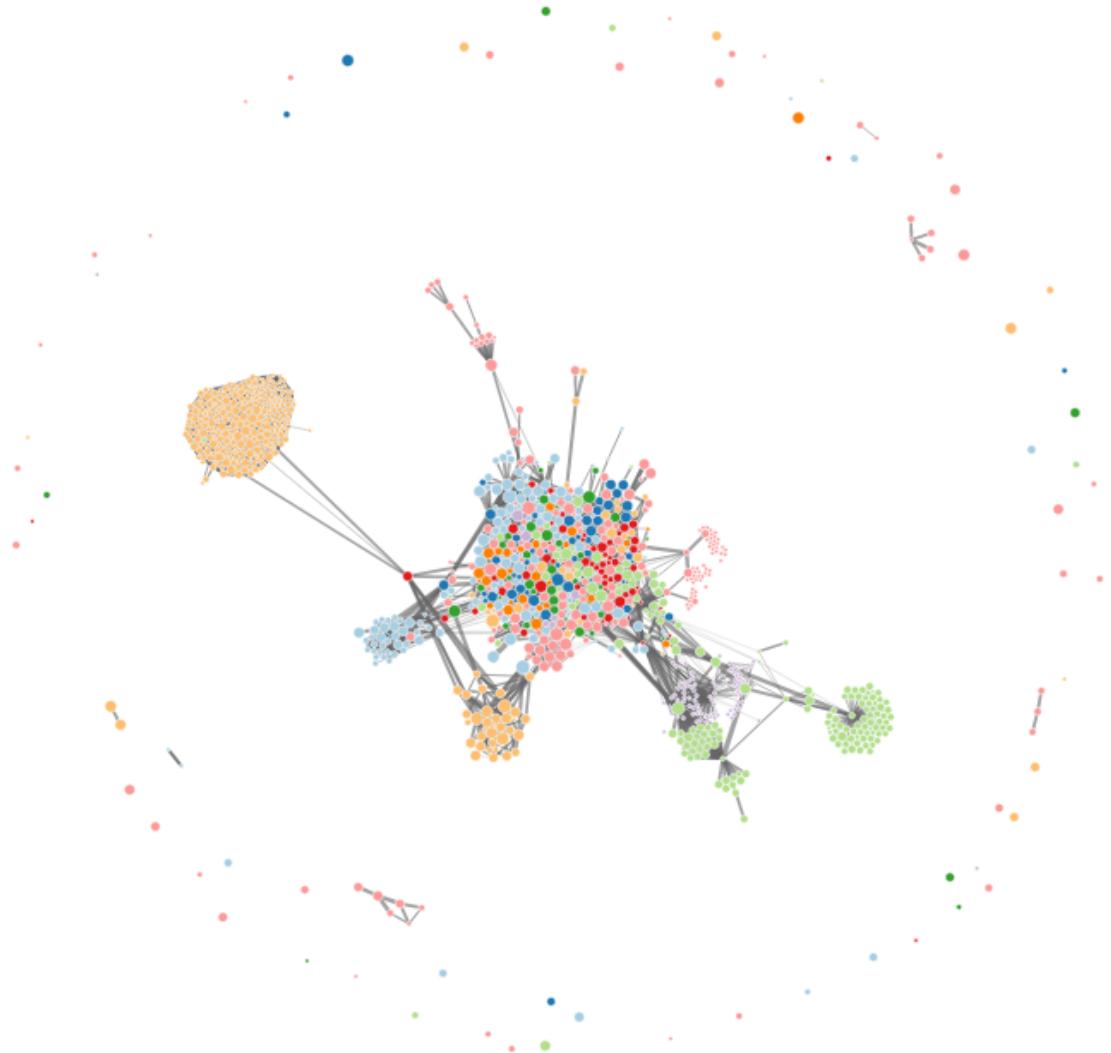


Figura 14: DAC - Raio do nó = $\log_{10}(\#\text{triplas})$

Nos pareceu importante ajustar o tamanho dos nós, se quisermos utilizar o raio do nó para codificar o seu número de triplas (importante comparar com o tamanho da LOD Cloud original, onde o tamanho do nó representa o número de enlaces de entrada e saída).

Depois destes ajustes no desenho, onde a parte central da nova imagem pode ser vista na Figura 15: DAC - Raio do nó = $\log_{10}(\#\text{triplas})$, temos uma representação mais adequada visualmente dos nós e seus tamanhos que a anterior, mas ainda será necessário ajustar mais parâmetros do diagrama de força, já que os nós sem enlaces ficaram muito dispersos e os clusters ainda estão distanciados demais, para que possamos alcançar uma representação mais satisfatória.

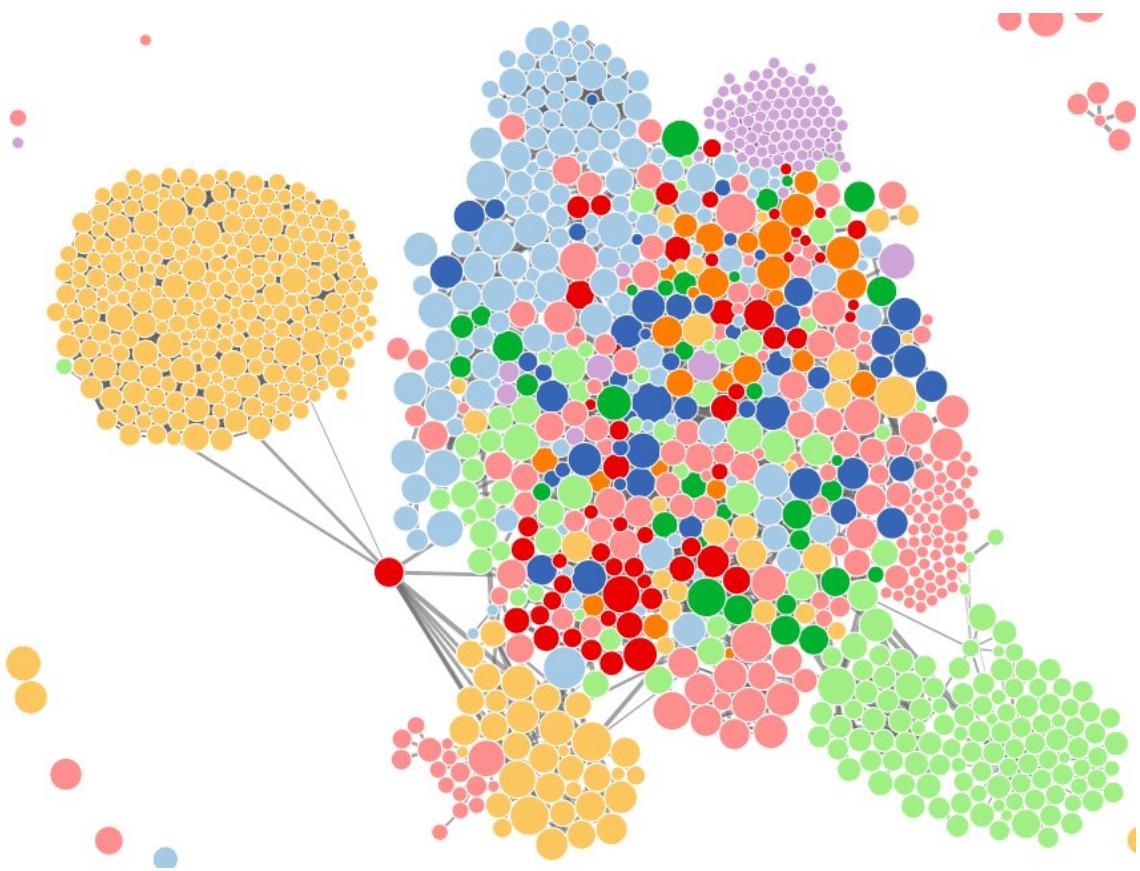


Figura 15: DAC - Raio do nó = $\log(\#\text{triplas})$

Com a representação mais compacta a informação de enlace se torna ainda mais irrelevante, por estar irremediavelmente escondida por trás dos nós. O recurso para mostrar esta informação, que em um desenho estático ficaria oculto, deverá incluir alguma forma de interatividade que permita visualizar os enlaces em separado dos nós, ou todos em conjunto mas com a opacidade dos nós atenuada.

Chegamos então à terceira tentativa de representação dos dados, mostrada integralmente na Figura 16: DAC - Raio do nó = $\log_2(\#\text{triplas})$.

Nesta última visualização preliminar todos os 1.163 nós da rede de Dados Abertos Conectados estão visíveis, o gráfico está compactado o suficiente para ficar visível em apenas uma tela e talvez seja possível visualizar os enlaces de forma independente com mais informação sobre o grafo. O raio dos nós (em pixels) é proporcional ao \log_2 do número de triplas, desta forma um nó com duas vezes mais triplas que outro menor, tem um pixel a mais de raio que o nó menor. A utilização de uma escala logarítmica para o tamanho do raio dos nós é importante porque há datasets com indicação de 0 tripla e outros datasets com mais de 9.500.000.000 triplas.

Antes de prosseguirmos, é importante frisar que estes datasets com 0 tripla (o que é um contrassenso sob o ponto de vista de dados abertos conectados) na verdade possuem uma quantidade não especificada de triplas, mas são tratados como se tivessem 0 tripla.

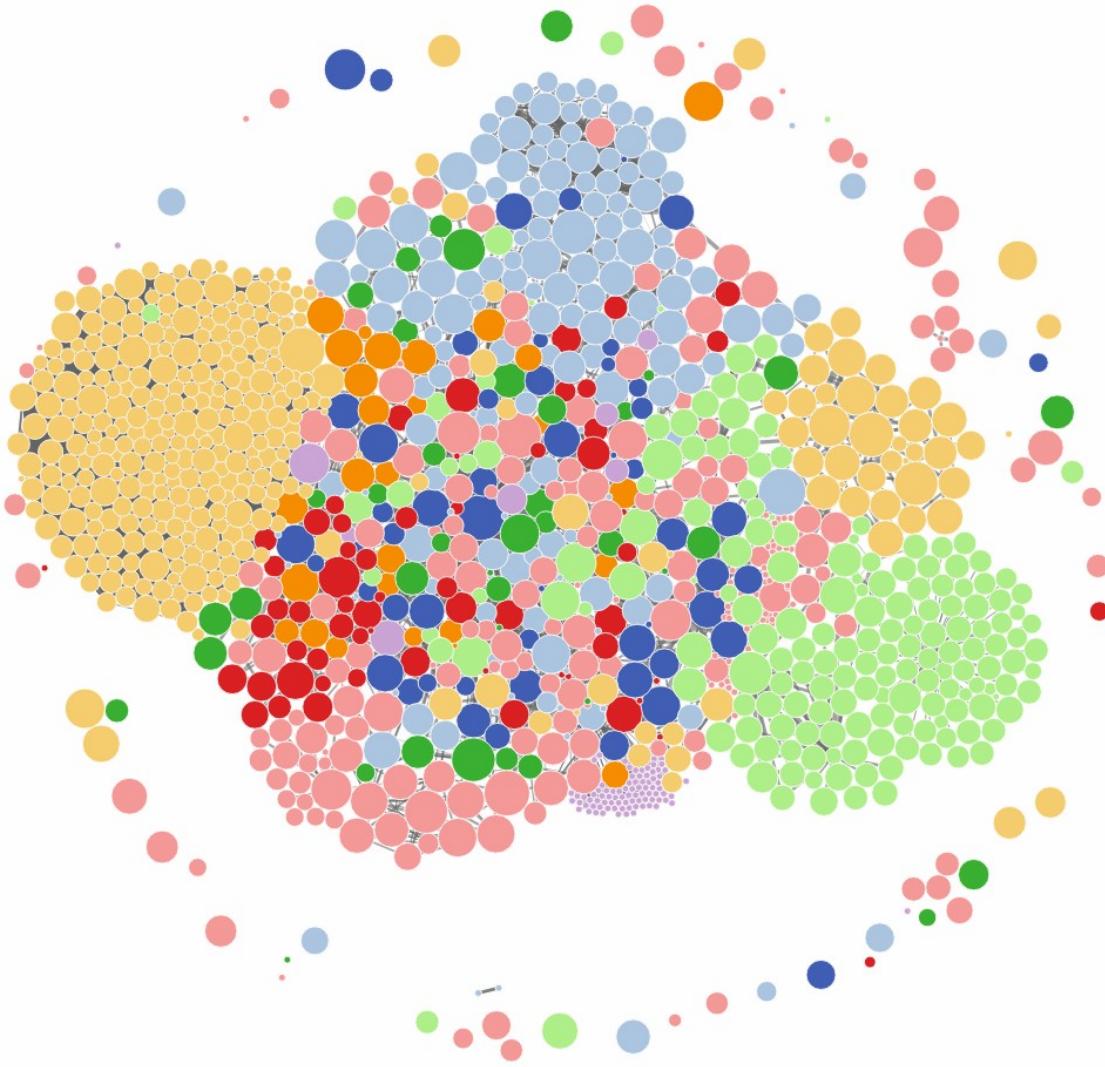


Figura 16: DAC - Raio do nó = $\log_2(\#\text{triplas})$

No gráfico da Figura 16: DAC - Raio do nó = $\log_2(\#\text{triplas})$, temos a informação de enlaces codificada mas coberta pelos próprios nós da rede. Para entender melhor os dados e buscar alternativas de desenho da versão final do gráfico que iremos propor, visualizaremos apenas os enlaces, sem a informação de nós.

A Figura 17: DAC - Apenas enlaces mostra a imagem resultante ao deixar os nós com total transparência. Nesse gráfico a espessura do enlace é proporcional ao \log_2 do número de triplas referenciadas, não há distinção de nós de entrada e saída (nós de entrada e saída para os mesmos pares de nós estão sobrepostos e apenas a conexão mais forte – com maior quantidade de triplas – está visível).

A análise deste último gráfico, em particular o clique formato pela categoria Life Sciences, mostra que não é possível derivar muita informação dos enlaces com a representação atual. Para tanto será necessário utilizar alternativas de interatividade e provavelmente utilizando um desenho diferente do mostrado na Figura 16: DAC - Raio do nó = $\log_2(\#\text{triplas})$.

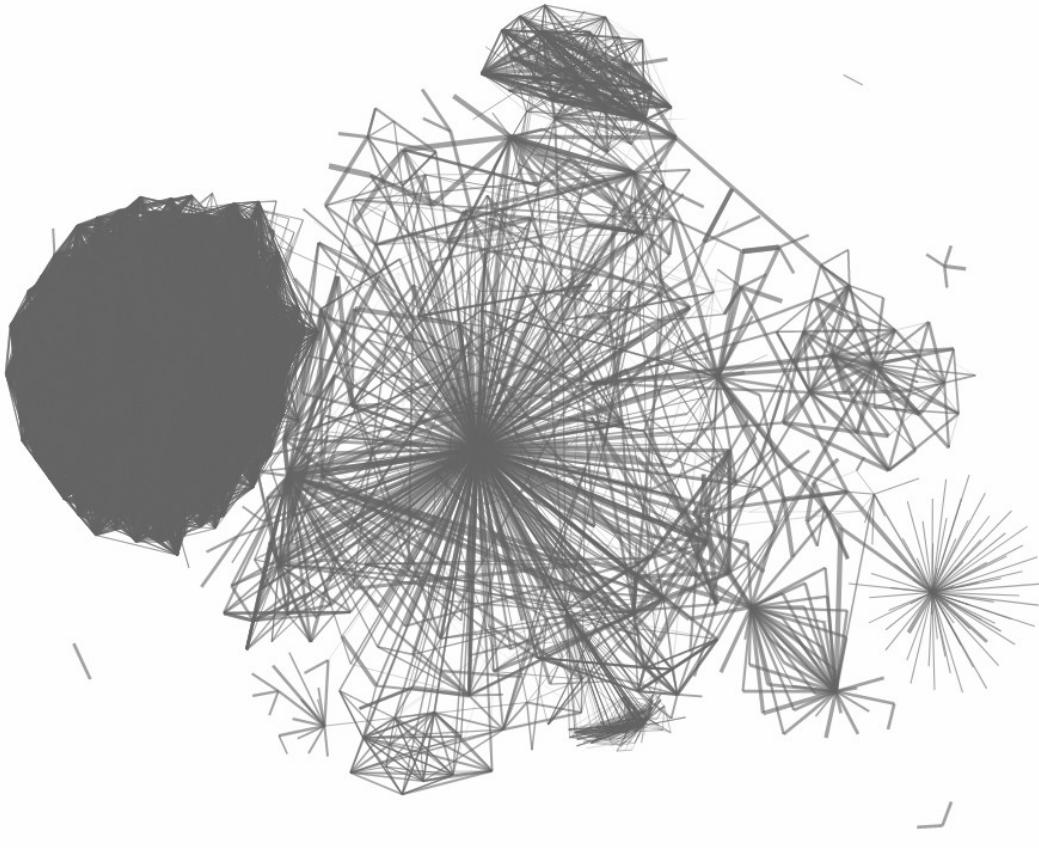


Figura 17: DAC - Apenas enlaces

5.5 Redesenho

As observações das seções anteriores ofereceram “insights” que nos permitiram chegar a uma proposta final de desenho de visualização para os datasets de Dados Abertos Conectados, que chamamos de LOD Target. Neste novo desenho pretendemos:

- Apresentar alternativas de visualizar os 1.163 nós atuais e as 15.655 arestas.
- Manter o raio dos nós proporcionais ao número de triplas do dataset, em uma escala log2.
- Cada uma das categorias será indicada por uma cor.
- Os nós manterão o máximo de informações do dataset dos arquivos originais.
- Os enlaces serão ocultáveis.
- Os nós (datasets) serão ocultáveis.
- O gráfico deve ser gerado de forma consistente (campo de força “controlado”).

Abaixo, na Figura 18: LOD Target - Primeira versão, mostramos a primeira versão do LOD Target, que atende a estes objetivos. Nesta imagem, gerada por uma página

JavaScript baseada no framework de visualização D3, temos todos os datasets da versão atual da LOD Cloud visíveis, com as categorias identificadas por cores em círculos concêntricos.

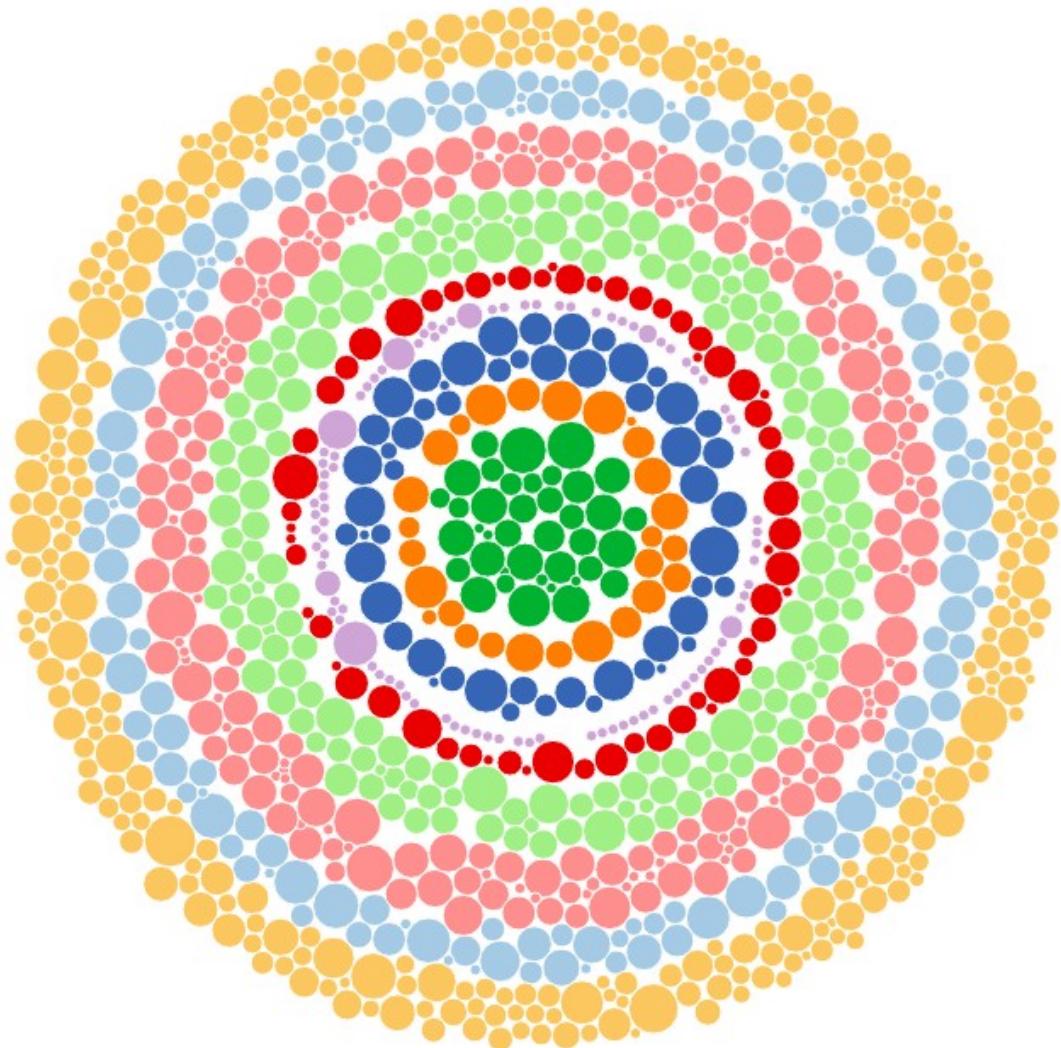


Figura 18: LOD Target - Primeira versão

As cores foram escolhidas utilizando-se de uma escala qualitativa de 9 classes do ColorBrewer 2.0. O ColorBrewer foi desenvolvido por Cynthia A. Brewer e Mark Harrower na Pennsylvania State University com o objetivo de ser uma referência com base científica para utilização de cores em mapas e gráficos.

A ordem das categorias nos círculos concêntricos foi escolhida de forma a manter a categoria de Life Sciences no círculo mais externo, assim mantendo uma máxima distância entre os vários nós deste clique hiperconectado, e as outras categorias ajustadas pelo número de datasets e área ocupada pelos círculos do dataset.

Uma das categorias, Social Networking, é degenerada no sentido de possuir um grande número de datasets sem informação do número de triplas. Na representação acima e na versão mais recente a categoria de Social Networking está representada como a quarta camada, de dentro para fora.

É possível alterar a opacidade dos enlaces, que estão transparentes na figura 18, e dos datasets, que estão totalmente opacos na figura 18, para explorar a conectividade da LOD Target. Na Figura 19: LOD Target - Enlaces, vemos apenas os enlaces com opacidade 0.05. Apesar da quantidade de linhas é possível explorar o gráfico e descobrir características dos enlaces e nós.

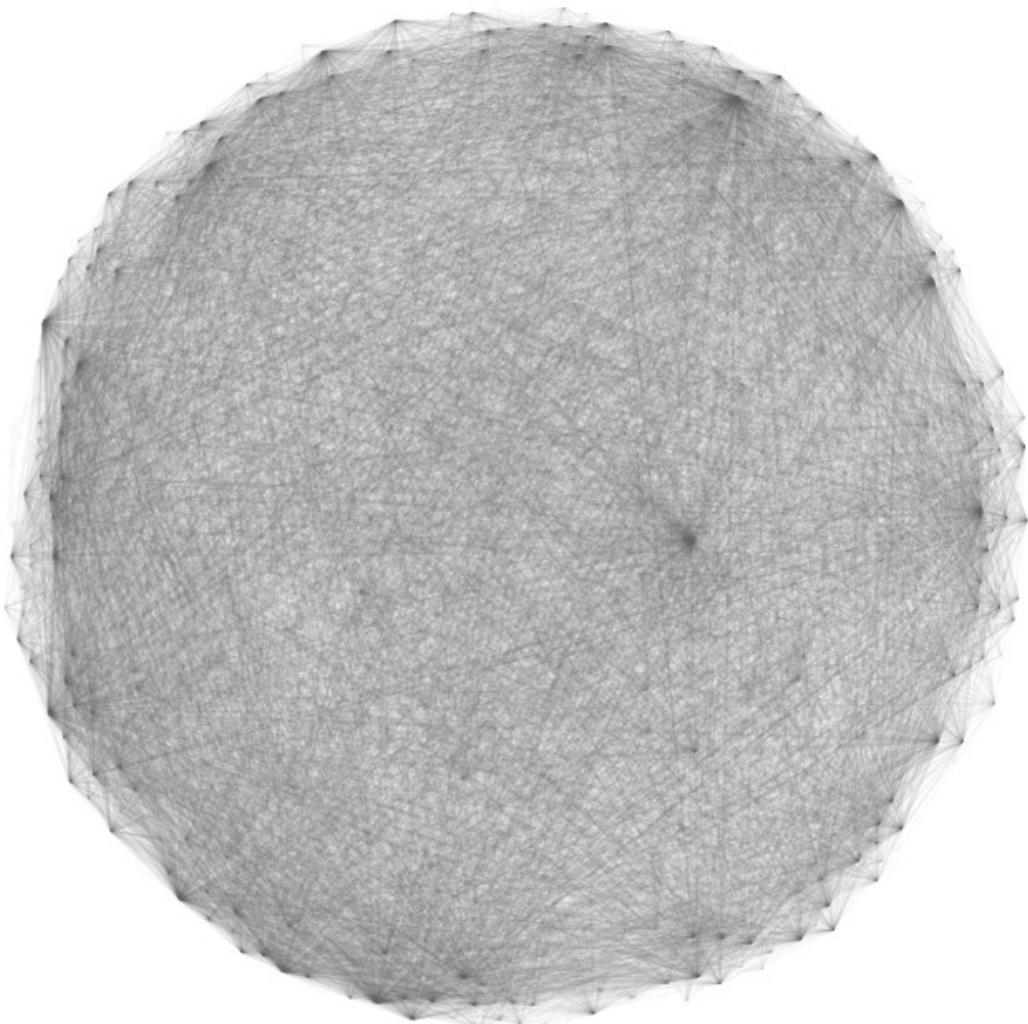


Figura 19: LOD Target - Enlaces

Na versão interativa os nós estão transparentes mas ainda presentes e ativos na imagem, assim, ao passar o mouse sobre o ponto mais escuro próximo ao centro da imagem vemos que este ponto altamente conectado é a DBpedia!

É possível também configurar a visualização de forma a diferenciar os nós e enlaces no momento que o mouse se sobrepõe a um dataset. Uma alternativa de visualização é mostrada na Figura 20: NIFSTD. Onde a opacidade de todos os outros nós foi atenuada e os enlaces do dataset NIFSTD foram destacados.

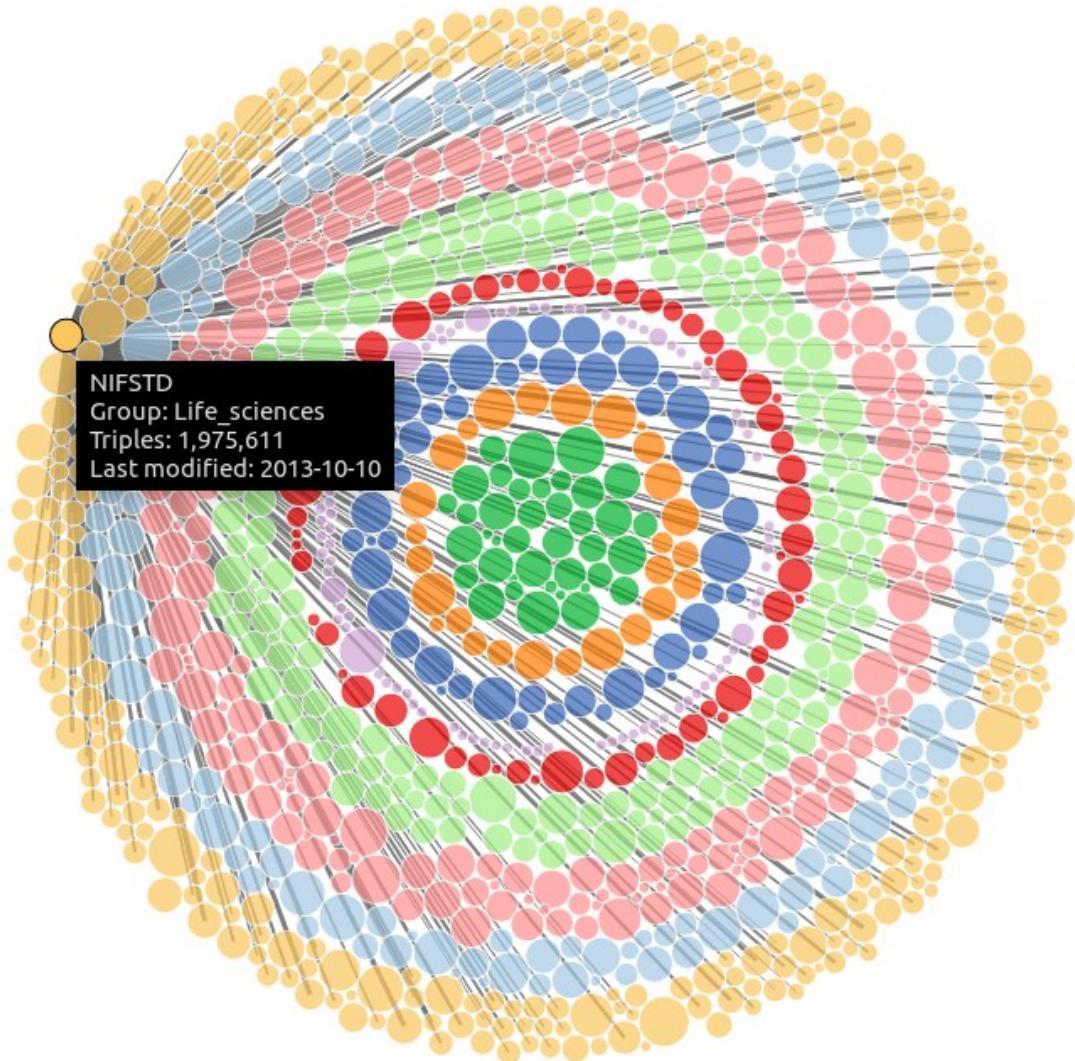


Figura 20: NIFSTD

Outra estratégia de visualização dos enlaces pode atenuar ainda mais os nós e ressaltar os enlaces de forma a destacar as conexões. A Figura 21: statusnet-piana-eu mostra esta outra versão, em que a ênfase sobre os enlaces é ainda mais evidente. Esta imagem já utiliza o modelo de cores da versão final da LOD Target,

Há várias outras alternativas de visualização e exploração que podem ser implementadas sem muita dificuldade no modelo atual da LOD Target, abaixo listamos algumas:

- Na exibição de enlaces mostrar apenas as conexões com mais que um certo número de triplas, ou com menos que um certo número de triplas, ou com um número de triplas dentro de um intervalo determinado.
- Separar na visualização as triplas de entrada das triplas de saída.
- Gerar pequenos múltiplos das várias visualizações de enlaces (ou seja, vários gráficos dos enlaces com cada um dos tamanhos determinados, exibidos lado a lado).

- Gerar pequenos múltiplos das várias categorias, como na Tabela 4.
- Esmaecer ou realçar apenas uma categoria determinada e suas conexões.
- Ao passar o mouse sobre um dataset, mostrar suas triplas de saída, ao clicar no dataset, mostrar as triplas de entrada.
- Visualizar apenas uma categoria e suas interconexões, reorganizando os nós se necessário.
- Gerar versões estáticas, em SVG compactado, de versões e subconjuntos específicos do LOD Target, para o desenvolvimento de narrativas específicas.
- Gerar uma versão compactada do gráfico em SVG sem a informação de enlaces e gerar as informações de enlaces em tempo real.

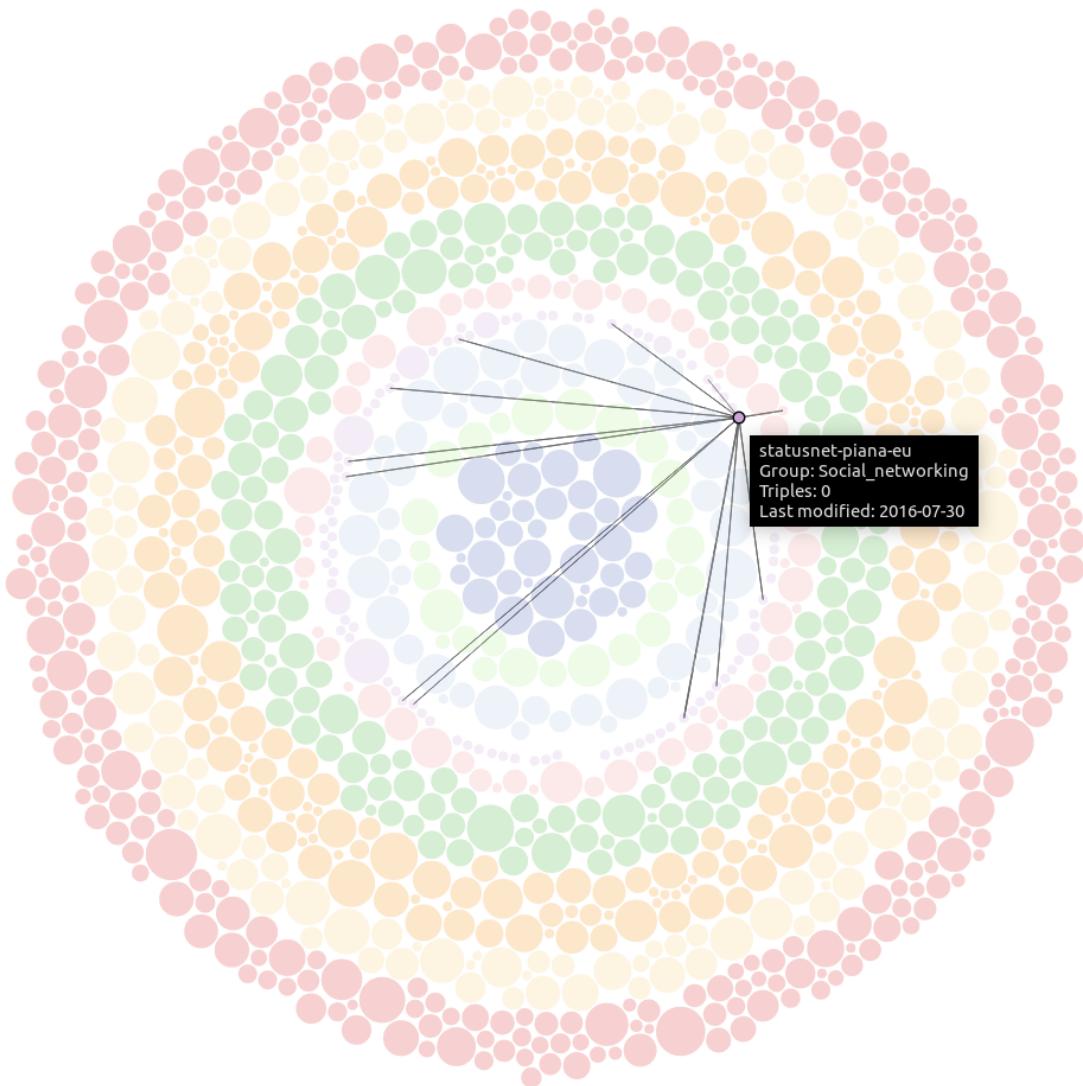


Figura 21: statusnet-piana-eu

Estas e outras estratégias podem facilmente ser desenvolvidas a partir da versão mais atual da LOD Target disponível no Github.

A Figura 22: LOD Target - Versão final mostra a última versão da LOD Target com a codificação de cores das categorias redefinidas para ficarem o mais próximo possível

das cores utilizadas pela LOD Cloud, mas ainda respeitando o modelo de cores qualitativo do ColorBrewer 2.0.

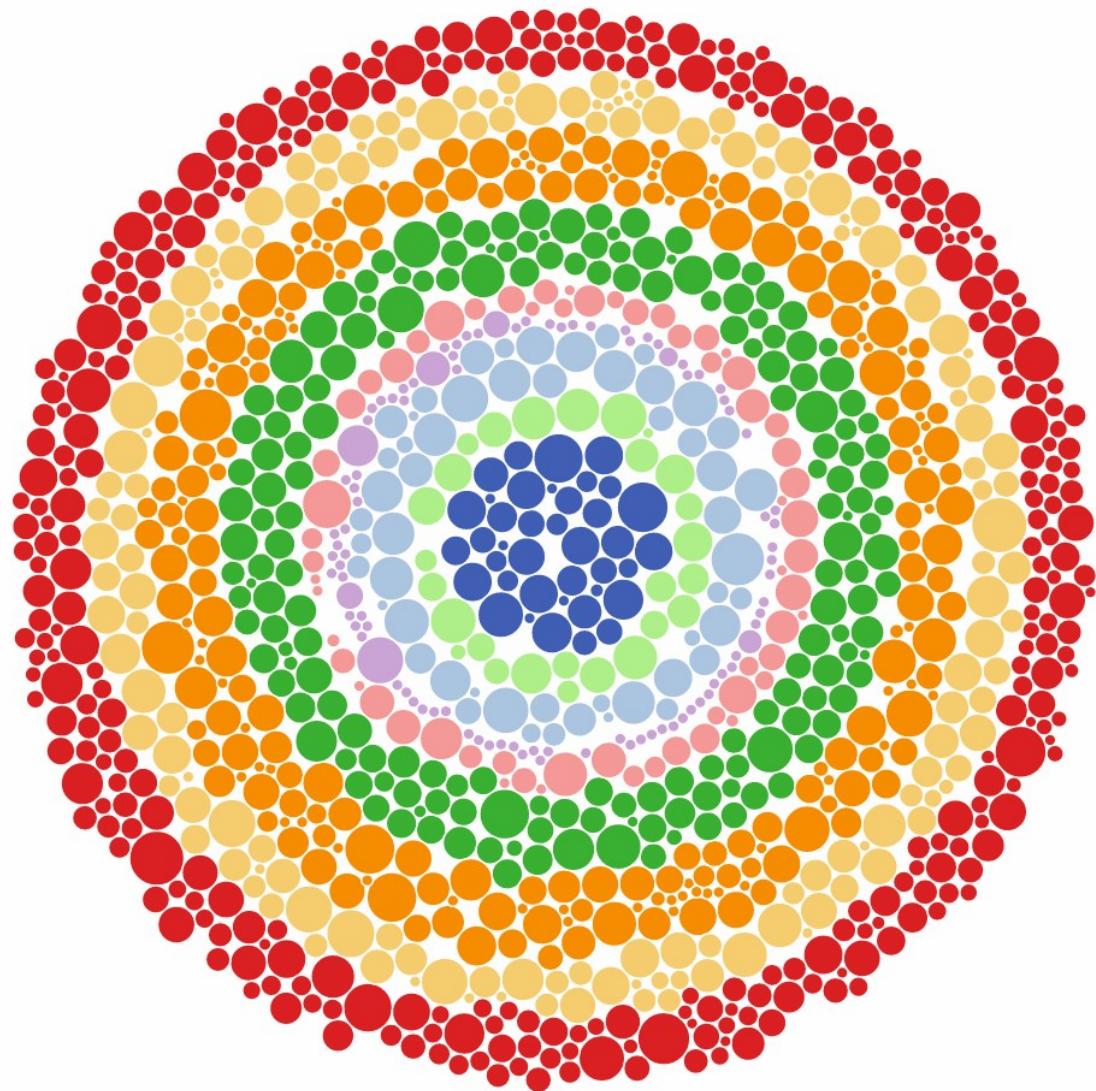


Figura 22: LOD Target - Versão final

Nesta versão as categorias estão indicadas pela legenda abaixo:

- [Red square] Life Sciences
- [Orange square] Publications
- [Yellow/orange square] Government
- [Green square] Linguistics
- [Pink square] User Generated
- [Purple square] Social Networking

 Cross Domain

 Media

 Geography

6. Github

O repositório do LOD Target no Github contém os seguintes arquivos html:

LODTargetGeneration.html: Gera o gráfico para ser baixado em SVG

LODTargetInline.html: Versão com o gráfico inline (na própria página HTML)

LODTargetInlineMin.html: Versão com gráfico inline, sem informação de enlaces

LODTargetWorker.html: Gera o gráfico com um Web Worker, que não retém o browser durante o processo de geração da imagem.

nodeSVG.html: Simulação do diagrama de força dos datasets

nodeSVG2ndTake.html: Outra versão da simulação do diagrama de força dos datasets

nodesExploration.html: Versão canvas do diagrama de força, sem informação de categorias, tamanho dos datasets e tamanho dos enlaces.

7. Limitações e futuro

A visualização LOD Target é efetiva porque a quantidade de datasets (1.163) ainda pode ser exibida em sua totalidade de forma coerente em uma página web e na tela do computador. Estimamos, porém, que a visualização não será mais adequada para visualizar TODOS os datasets ao mesmo tempo quando este número ultrapassar a quantidade de 2.000 datasets. A partir desta quantidade de datasets, será necessário, no mínimo, ajustar a visualização e talvez seja necessário criar um novo desenho para a visualização dos datasets. Uma estratégia que pode exigir menos esforços de adaptação, seria manter a mesma estrutura da visualização atual, mas incluindo uma camada hierárquica. Nesse caso há várias alternativas para hierarquia, mas a mais natural é a separação por categorias, onde poderíamos ver as conexões entre as diversas categorias no nível mais alto e internamente em cada categoria nos níveis mais baixos.

O crescimento do número total de enlaces e de suas densidades não será um problema enquanto o total de datasets se mantiver em um tamanho controlado, já que o pressuposto é que os enlaces são visualizados de forma interativa e com filtros que permitem a seleção de enlaces de forma coerente com o objetivo da investigação dos dados.

Outras funcionalidades e definições podem ser incluídas em versões futuras da LOD Target, por exemplo:

- Incluir anotações nas visualizações
- Incluir busca por dataset
- Usar mais de uma representação para a nuvem. Target+Matriz. Incluir Hive plot.
- Incluir critérios mais definidos de escalabilidade
- Formalizar hierarquia grupo→nós

- Definir vocabulário da nuvem LOD
- Definir classificação de grupos da nuvem LOD
- Triplificar a nuvem LOD

Referências

- Abele, A., McCrae, J. P., Buitelaar, P., Jentzsch, A., Cyganiak R. Linking Open Data cloud diagram 2017. Disponível: <http://lod-cloud.net/>. Acesso novembro/2017
- Berners-Lee, T. “Linked Data”, Disponível: <https://www.w3.org/DesignIssues/LinkedData.html>, Julho/2006. Acesso: novembro/2017.
- ColorBrewer 2.0. Disponível: <http://colorbrewer2.org/>. Acesso: novembro/2017.
- Data Hub. Disponível: <http://datahub.io/>. Acesso: novembro/2017.
- Force Directed Graph. Disponível: <https://bl.ocks.org/mbostock/4062045>. Acesso: novembro 2017.
- Interactive Cloud Image 2017-01-26. Disponível: <http://lod-cloud.net/versions/2017-01-26/cloudImage2017.svg>. Janeiro/2017. Acesso: novembro/2017.
- Les Misérables Co-occurrence. Disponível: <https://bost.ocks.org/mike/miserables/>. Acesso: novembro/2017
- Linking Open Data. Disponível: <https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>. Acesso: novembro/2017.
- LOD Target. Disponível: <https://github.com/RobStelling/LODChart>. Acesso: novembro/2017.
- Munzner, T. Visualization Analysis and Design (AK Peters Visualization Series), A K Peters/CRC Press, 1^a edição, Dezembro 2014.
- The Linking Open Data cloud diagram. Disponível: <http://lod-cloud.net/>
- World Internet Topology. Disponível: http://www.research.att.com/export/sites/att_labs/groups/infovis/news/img/ATT_Labs_InternetMap_0730_10.pdf. Acesso: novembro/2017