ADOÇÃO DO XBRL EM PROJETOS OPEN SOURCE: MINERAÇÃO NO REPOSITÓRIO GITHUB

Aloisio Sampaio Cairo, Diego Simões Santana, Leonardo Sampaio Cairo, Paulo Caetano da Silva
Universidade Salvador (UNIFACS), Brazil
aloisiocairo@gmail.com, diegosimoessantana@gmail.com, leocairos@gmail.com, paulo.caetano@pro.unifacs.br

Resumo: **XML por ser uma linguagem estruturada, com fácil visualização e entendimento, tanto para softwares como para as pessoas, possibilitando que seja utilizada em qualquer projeto, permite que ocorra uma interação de diferentes softwares, com a criação e leituras de arquivos. Baseado em XML, porém, direcionada para a divulgação de demonstrativos financeiros, foi desenvolvida a linguagem XBRL com o propósito de possibilitar a criação de regras nos documentos, reduzindo a ambiguidade, definindo os conceitos de metadados, por meio de linkbases. Com características e validações para o ramo financeiro, XBRL tem sido adotada em grandes projetos de softwares Open Source. Por isto, este artigo faz uma pesquisa sobre o nível de adoção do XBRL em projetos hospedados no repositório GitHub e com isso além da quantidade de projetos que adotam XBRL, serão apresentadas também as características destes projetos.**

Palavras chave: **XML; XBRL; GitHub; Projetos XBRL**

I. INTRODUÇÃO

Com o grande número de informações criadas no dia a dia das organizações e com a necessidade dessas informações serem armazenadas em locais seguros e que estejam disponíveis a qualquer momento, surge a necessidade de organizações possuírem um software que as manipule. Os dados podem ser processados, gerando relatórios que auxiliem na tomada de decisões, assim como integrá-los a outros softwares externos com características especificas, podendo essa integração ser realizada através de arquivos XML.

XML apresenta diversas características: é utilizável na internet, legível por pessoas, possibilita a publicação eletrônica de documentos por um meio independente, permite a troca de informações independentemente da plataforma de hardware e software [10]. Apresentando todas essas características e obedecendo uma estrutura e hierarquia, XML possibilita que os dados sejam lidos, interpretados e integrados por pessoas ou software. Herdando características da XML, a XBRL, permite controle sobre as informações financeiras da organização aumentando a eficiência e eficácia nas decisões com o objetivo de se tornar a linguagem padrão para a divulgação de dados financeiros.

Este artigo apresenta características e aspectos gerais de XML, XBRL e do repositório GitHub. São apresentados dados quantitativos de projetos Open Source que são armazenados no GitHub e que utilizam XBRL. Esses dados são apresentados em formato gráfico e explicados separadamente, por fim, conclui-se discutindo a adoção de XBRL por projetos de software e propõe-se alguns trabalhos futuros. Desta forma, este artigo está organizado da seguinte forma, na Seção II discute-se o referencial teórico necessário ao seu entendimento, na Seção III é apresentada a utilização do XBRL em projetos Open Source e métodos aplicados para identificar o nível de adoção do XBRL, na Seção IV conclui-se se o XBRL está sendo incluso em projetos, assim como discute-se os trabalhos futuros.

II. REFERENCIAL TEÓRICO

Para um bom entendimento deste trabalho é necessário que alguns conceitos sejam introduzidos, sobre XML, XBRL e o repositório GitHub, por fim, algumas práticas sobre mineração nesse repositório são discutidas.

II.I. XML

XML (eXtended Markup Language), é um padrão de troca de informações recomendado pela organização de padronização W3C (Word Wide Web). Arquivos XML possuem uma estrutura de fácil entendimento e é de fácil interpretação, podendo cada arquivo ter diferentes tipos de tags. XML possui diversos benefícios para desenvolvedores e usuários, sendo eles:

- O documento XML tem uma lógica e uma estrutura física. Sua estrutura física é composta por entidades. Entidades se relacionam com outras entidades. O documento inicia-se com uma entidade "raiz", é composto por declarações de elementos, comentários, referências e outros tipos de componentes. [10]
- De acordo a W3C, os principais objetivos estabelecidos na especificação da linguagem XML se caracterizam por: ser utilizável na Internet; ser possível fazer a leitura por pessoas; possibilitar publicação eletrônica; definir protocolos para troca de dados; possibilitar que o arquivo seja processado por software de baixo custo; facilitar a utilização de metadados e auxiliar na recuperação dos dados na internet.

Com XML é possível criar documentos com dados organizados e obedecendo a hierarquias. XML não depende de hardwares ou de softwares, por isso qualquer aplicação pode criar um arquivo XML e esse arquivo ser lido por qualquer outra aplicação.

II.II. XBRL

XBRL (eXtensible Business Reporting Language) é um padrão internacional baseado em XML e mantido por um consórcio internacional sem fins lucrativos (XBRL International Incorporated). Este padrão é utilizado por mais de 50 países para emissão de relatórios financeiros em formato digital, por serem mais eficazes, organizadas e precisos [12].

Algumas características da XBRL são [12]:

- **Definições claras:** Permite criar taxonomias que captura o significado e a relação entre os termos contido em relatórios. As taxonomias são elaboradas por reguladores, agências governamentais, ou qualquer organização que queira intercambiar dados com base em XBRL. Os tipos de informações são ilimitados, podendo ser estendida, para cada necessidade do usuário dos documentos XBRL;
- **Regras de negócio testável:** XBRL permite que sejam criadas regras lógicas ou matemáticas que restringe o

que pode ser publicado ou intercambiado. As regras criadas podem ser utilizadas para: identificar ou destacar informações que podem ser questionadas; criar índices ou informações que agreguem valor.

- **Suporte multilíngue:** XBRL permite que as definições dos conceitos sejam elaboradas em diversos idiomas, permitindo criar relatórios financeiros em diferentes linguagens naturais.
- **Suporte a software:** XBRL é suportado por uma grande quantidade de software.

XBRL é utilizado por grandes entidades, reguladores de seguros, reguladores tributários e do mercado financeiro e bancário, empresas, pelo governo para informações intragovernamental, bancos de investimentos, bolsa de valores [7].

II.III. GITHUB

Atualmente os projetos de software possuem mecanismos de controle de versão. Por muito tempo foi usado o CVS[1] (Concurrent Version System) para realizar este controle. O SVN[2] (Apache Subversion) foi desenvolvido para ajustar algumas falhas do CVS e acrescentar novas funcionalidades. Uma alternativa para esses software de controle de versão é o GitHub

O Git é um sistema de controle de versão de arquivos através do qual é possível desenvolver projetos nos quais diversas pessoas podem contribuir simultaneamente, editando e criando novos arquivos. Desta forma, o Git permite que os arquivos do projeto possam existir sem o risco de terem suas alterações sobrescritas. E, além disso, outro fator importante do Git (um dos seus diferenciais em relação ao SVN) é a possibilidade de criar, a qualquer momento, vários snapshots (branch) do projeto.

O GitHub é um serviço de hospedagem na internet para compartilhamento de projetos (públicos e privados) usando o controle de versionamento Git. No GitHub é possível seguir outros desenvolvedores, baixar projetos, modificar projetos, receber atualizações de modificações de projetos, entre outras funcionalidades que o diferencia do SVN e CVS. Em seu repositório existem diversos projetos grandes que qualquer pessoa interessada pode visualizar os códigos fonte. Dentre esses projetos pode-se citar o JQuery, Eclipse, VRaptor, Twitter. Além disso, o Github possui funcionalidades de uma rede social como feeds, followers, wiki e um gráfico que mostra como os desenvolvedores trabalham nas versões de seus repositórios.

O Git se integra com o GitHub de forma bem simples. E com isso, é possível criar um repositório no GitHub e simplesmente "commitar[3]" as alterações do projeto Git local, tornando-o público (com uma conta paga é possível também "commitar" projetos tornando-os privado).

> "Milhões de desenvolvedores usam o GitHub para construir projetos pessoais, apoiar seus negócios e trabalhar juntos em tecnologias open source. (...). Estamos apoiando uma comunidade onde mais de 15 milhões de pessoas aprendem, compartilham e trabalham juntos para construir software" [3]

II.IV. MINERAÇÃO de DADOS no GITHUB

Mineração de dados é uma técnica utilizada para explorar uma grande quantidade de dados com o intuito de criar associações e padrões transformando os dados em informações. Pode ser dividida em dois tipos: direcionada e não direcionada. O tipo direcionado tem o objetivo de tentar identificar um aspecto em particular, como por exemplo: o preço de um carro baseado em outros carros da mesma categoria e do mesmo ano; e o tipo não direcionado que tenta encontrar padrões em dados que já existem [6].

O GitHub fornece uma API para acesso a sua base de dados de projetos. Com esta API é possível extrair diversas informações dos projetos hospedados. A API de busca é otimizada para encontrar o item específico que está sendo procurado (e.g. um usuário específico, um arquivo específico em um repositório). Para satisfazer essa necessidade, o GitHub Search API fornece até 1.000 resultados para cada pesquisa.

A forma mais simples de utilizar esta API é através do CURL[4]. A seguir é ilustrado um exemplo de uso da API para listar todos os projetos no GitHub com a chave de busca "XBRL":

curl 8 Estudo da Integração de um Array Adaptável em Processadores Multicore.doc

O resultado da pesquisa é exibido em formato JSON, como ilustrado na Figura 1.

[1] http://savannah.nongnu.org/projects/cvs/
[2] https://subversion.apache.org/
[3] Salvar uma alteração de código-fonte no sistema Git de versionamento do Github

[4] Ferramenta de código aberto e uma biblioteca para transferência de dados com a sintaxe de URL. É utilizada em linhas de comando ou scripts para transferir dados.

```
{
  "total_count": 157,
  "incomplete_results": false,
  "items": [
    {
      "id": 23014857,
      "name": "xbrl",
      "full_name": "adrianclay/xbrl",
      "owner": {
        "login": "adrianclay",
        ...
      },
      ...
      "description": "A simple API to parse XBRL taxonomies in PHP.",
      ...
      "created_at": "2014-08-16T09:23:11Z",
      "updated_at": "2016-05-31T14:05:09Z",
      ...
      "size": 355,
      ...
      "watchers_count": 8,
      "language": "HTML",
      ...
      "score": 43.669563
    },
```


Figura 1 – Arquivo JSON com resultado do CURL

III. XBRL EM PROJETOS OPEN SOURCE

O projeto GHTorent disponibiliza uma cópia off-line, escalável e consultável dos dados do GitHub, dados que são acessíveis através da API do GitHub [5]. Na página do projeto consta que o GHTorent já possui uma base com mais de 6,5 TB de dados brutos em formato JSON, e mais de 40 GB de dump do MySQL.

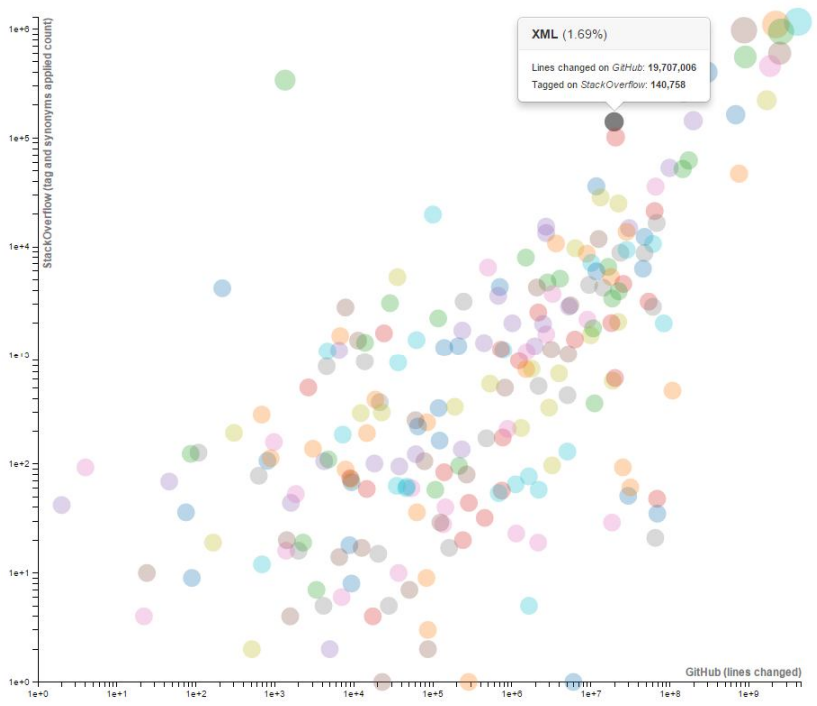
O gráfico mostrado na Figura 2, apresenta a popularidade das linguagens no GitHub e no StackOverflow[5].

De acordo com o [4]:

"Os dados do GitHub exibidos no gráfico, Figura 2, baseiam-se nos resultados on-line da pesquisa dos eventos da API do GitHub. Eles são atualizados instantemente sempre que é identificado um novo evento de push no GitHub. Os dados do StackOverflow são baseados no número de vezes que uma marca para uma determinada linguagem é aplicada, junto com a contagem dos sinônimos dessa língua (atualizados a cada quatro horas). A porcentagem que é mostrada é relação da média GitHub e StackOverflow para uma linguagem".

Nota-se que XML está entre as 20 linguagens mais utilizadas. Apesar de XBRL não aparecer no gráfico, o GitHub conta hoje com diversos projetos XBRL, os quais estão contidos nos projetos do tipo XML.

Neste trabalho será apresentado dados a respeito da



JavaScript	Java	C#
PHP	Python	HTML
C++	CSS	SQL
Objective-C	C	Ruby
R	XML	Swift
Matlab	Scala	Perl
Shell	Delphi	PowerShell
Haskell	XSLT	Assembly
Nginx	Go	Awk
Groovy	Makefile	PLSQL
TypeScript	Clojure	Lua
ColdFusion	F#	CoffeeScript
ActionScript	Cuda	Arduino
CMake	Prolog	SAS
Erlang	FORTRAN	Dart
Visual Basic	Cucumber	AppleScript
Scheme	Rust	GLSL
Tcl	OCaml	QML
XPages	Max	Processing
Smarty	Gnuplot	VHDL
Diff	LLVM	XQuery
Verilog	ANTLR	Puppet
Racket	AspectJ	Elixir
Pascal	FreeMarker	PLpgSQL
D	Mask	Common Lisp
Emacs Lisp	Thrift	Julia
NSIS	AutoIt	RobotFramework
Bison	NetLogo	AutoHotkey
Stata	Apex	Objective-C++
Handlebars	IDL	TeX

Figura 2 - Programming Language Popularity Chart (fonte: GHTorent)

[5] Website americano de perguntas e respostas sobre desenvolvimento de software com grande audiência no exterior.

adoção de XBRL em projetos Open Source registrados no repositório GitHub. Para tal, foi utilizada a API do GitHub por meio da ferramenta CURL.

Em razão das limitações impostas pela API, foi necessário realizar a paginação dos resultados da consulta para que fosse possível obter todos os projetos. Em um primeiro momento foi identificado que o GitHub possuía 157 projetos com referência ao XBRL. E em seguida foram realizadas 2 consultas para extração de dados destes projetos:

```
# curl
"https://api.github.com/search/repositories?q=xbrl&per_page=100&page=1"
>> xbrl_1_100.txt
```

```
# curl
"https://api.github.com/search/repositories?q=xbrl&per_page=100&page=2"
>> xbrl_101_157.txt
```

Cada um destes comandos possui capacidade para exibir resultados de até 100 projetos. Por isso, foi necessário mesclá-los para obtenção de resultado completo. Os dados foram gerados em formato JSON pela ferramenta CURL e em seguida exportados para o formato CSV em planilha eletrônica.

De acordo com os dados obtidos, o primeiro projeto XBRL foi criado em março de 2009 (Projeto mixanalytics de theRocket: Corporate data brought to your mobile phone via XBRL) e o ultimo em maio de 2016 (Projeto xbrl de Jeanferly: sem descrição).

Figura 3 – Gráfico anual de evolução de novos projetos

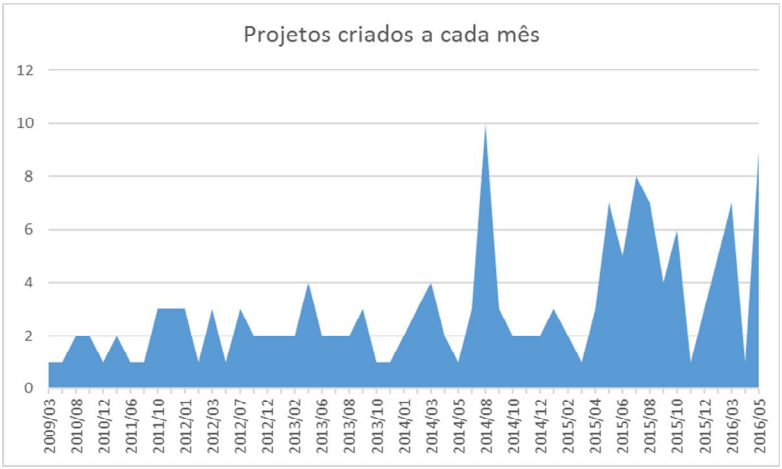


Figura 4 – Gráfico mensal de evolução de novos projetos

A seguir, na Figura 3 e 4, é possível observar a evolução anual e mensal da criação de projetos XBRL no GitHub.

Nota-se que houve uma forte oscilação a partir de 2014, quando o número de projetos criados a cada mês chegou a 10 repositórios. E a partir de 2015 foram criados em média 6 novos projetos a cada mês.

No gráfico anual, ilustrado pela Figura 4, fica mais clara a percepção de que a criação de projetos XBRL tem evoluído com o passar dos anos, tendo seu ápice em 2015. Com isso, espera-se que em 2016 esta evolução continue repercutindo positivamente, tendo em vista a quantidade atual de projetos criados neste ano até o presente momento.

Os estudos dos projetos XBRL no GitHub mostram que, dos 157 projetos localizados, 67 (43% do total) estão aparentemente abandonados, uma vez que não sofrem atualizações há mais de 12 meses. Destes 67 projetos, 29 não sofrem atualizações há mais de 24 meses e 9 há mais de 36 meses. Os demais projetos, 90 ao todo (57% do total), possuem uma média de atualização a cada 5 meses, a Figura 5 mostra essas informações.

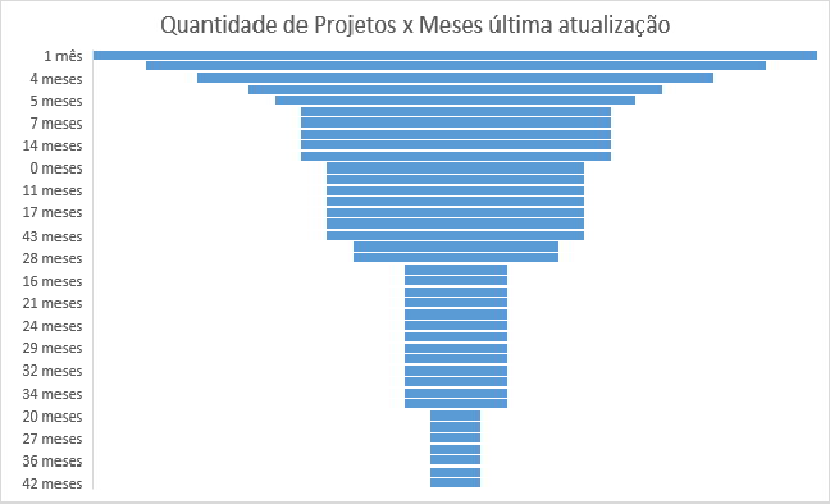


Figura 5 – Quantidade de Projetos x Meses última atualização

Os gráficos ilustram que o surgimento de novos projetos é proporcional a atualização destes e dos novos projetos, Figura 6. Ou seja, os projetos XBRL se mantêm ativos e atualizados no repositório do GitHub, Figuras 6 e 7. Entre aqueles projetos atualizados recentemente, destacam-se os projetos mostrados na Tabela 1.

Figura 6 – Projetos atualizados (mês)

Figura 7 –Projetos atualizados (ano)

Projeto	*Descrição*
XBRL	:exclamation: This is a read-only mirror of the CRAN R package repository. XBRL — Extraction of Business Financial Information from 'XBRL' Documents
pystock-crawler	Crawl and parse financial reports (XBRL) from SEC EDGAR, and daily stock prices from Yahoo Finance
Pysec	Parse XBRL filings from the SEC's EDGAR in Python
Litexbrl	XBRL parser for Ruby
python-xbrl	xrbl parser written in Python
ScraXBRL	SEC Edgar Scraper and XBRL Parser/Renderer
financial_fundamentals	Find XBRL filings on the SEC's Edgar and extract accounting metrics.
parse-xbrl	Parse xbrl documents to extract tags for common financial data
Xbrlparser	An incomplete XBRL library (in Ruby) for reading XBRL instance and taxonomy documents.
sec-xbrl	XBRL.US Webinar: How to download and process SEC XBRL Data Directly from EDGAR
XBRLFiles	Explore XBRL with R
xbrl2rdf	Publishing XBRL document as RDF data
xbrlware-ruby19	A re-packaging and ruby19-ification of xbrlware -- for pulling XBRL financial filings from SEC Edgar.

Tabela 1 - Projetos atualizados recentemente

Além disso, cabe destacar que esses projetos possuem até 117 desenvolvedores (observadores), mas em média, cada projeto possui sete observadores (mediana = 1), conforme ilustrado na Figura 7. Os projetos com o maior número de observadores são mostrados na Tabela 2.

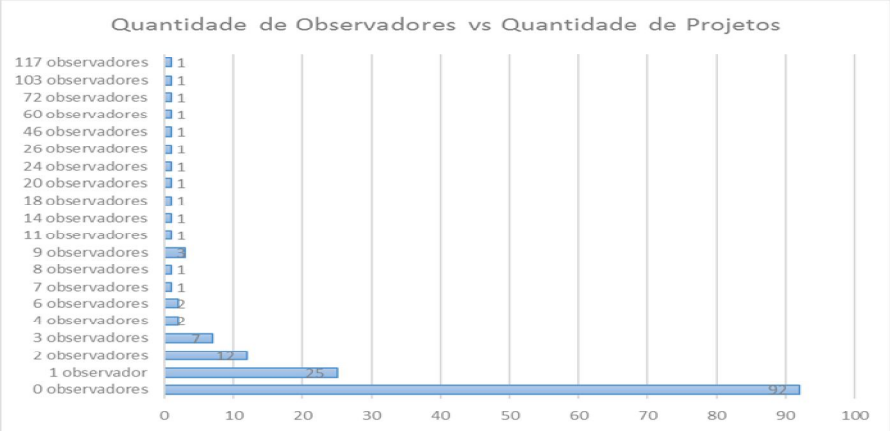


Figura 8 – Quantidade de Observadores vs Quantidade de Projetos

Proprietário	Projeto	Descrição do Projeto	Criado em	Atualizado em	Nº Observ.
lukerosiak (User)	pysec	Parse XBRL filings from the SEC's EDGAR in Python	04/05/2013	26/06/2016	117
Arelle (Organization)	Arelle	Arelle open source XBRL platform	08/06/2011	12/05/2016	103
eliangcs (User)	pystock-crawler	Crawl and parse financial reports (XBRL) from SEC EDGAR, and daily stock prices from Yahoo Finance	31/08/2013	01/07/2016	72
andrewkittredge (User)	financial_fundamentals	Find XBRL filings on the SEC's Edgar and extract accounting metrics.	27/02/2013	10/06/2016	60
greedo (User)	python-xbrl	xrbl parser written in Python	04/08/2014	23/06/2016	46
altova (Organization)	sec-xbrl	XBRL.US Webinar: How to download and process SEC XBRL Data Directly from EDGAR	20/09/2014	07/06/2016	26
computerpencils (User)	ScraXBRL	SEC Edgar Scraper and XBRL Parser/Renderer	06/03/2016	15/06/2016	24
bergant (User)	XBRLFiles	Explore XBRL with R	22/01/2015	05/06/2016	20
Arelle (Organization)	EdgarRenderer	EDGAR Renderer enables investors to view the interactive data filings submitted under the US Security and Exchange Commission (SEC) rules that require the use of XBRL via the SEC website.	17/04/2015	24/05/2016	18
Eddolan (User)	XBRL.js		13/05/2015	31/05/2016	14
MarkGannon (User)	XBRL	Perl Module for Reading XBRL	08/03/2012	01/01/2016	11

Tabela 2 - Projetos com o maior número de observadores

Daqueles projetos mostrados na Tabela 2, 20 projetos são de empresas/organizações e os demais 137 são de indivíduos, ou seja, a grande maioria dos projetos XBRL encontrados no GitHub são de cunho pessoal, Figura 9.

Figura 9 – Proprietários dos projetos

Os projetos encontrados, utilizam 19 linguagens de programação diferentes, sendo que, as mais representativas foram: Python (33), Java (20), Ruby (17) e JavaScript (13), a Figura 10 mostra um gráfico com a distribuição das linguagens usadas nos projetos.

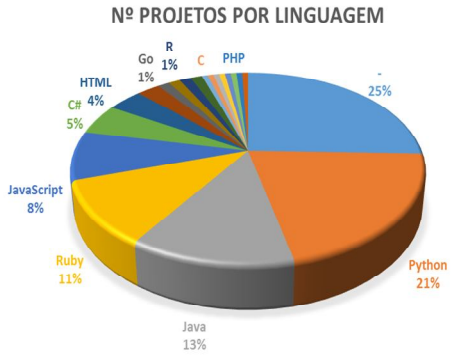


Figura 10 – Gráfico linguagens de programação em projetos XBRL

Por fim, foram contabilizadas as palavras mais utilizadas nas descrições dos projetos. Na Figura 11 é possível observar os termos mais utilizados nos projetos. Sua representatividade é proporcional ao tamanho em pixel de cada palavra abaixo em relação às demais.

Figura 11 – Nuvem de palavras mais utilizadas na descrição dos projetos XBRL

IV. CONCLUSÃO

A partir dos dados minerados no repositório do GitHub e apresentados neste artigo, projetos XBRL têm sido incluídos mensalmente no repositório Github como novos projetos Open Source e que mais da metade dos projetos são atualizados em média duas vezes ao ano, com projetos direcionados para empresas e para interesses individuais, sendo esses projetos desenvolvidos em linguagens de programação para a web, desktop e mobile. Portanto, a partir dos dados coletados e apresentados, pode-se concluir que XBRL está sendo utilizado por diversas empresas e interesses, possibilitando uma melhor qualidade da transparência e organização da saúde financeira de uma empresa.

Como trabalhos futuros, sugere-se uma pesquisa sobre a adoção de XBRL em projetos que são armazenados em outros repositórios, como por exemplo o Google Code, BitBucket, Assembia, Team foundation, estre outros, assim como a adoção em projetos proprietários. Desta forma poderá se ter um melhor entendimento acerca da adoção de XBRL e dos seus diversos projetos.

REFERÊNCIAS

[1] CURL. **Curl is an open source command line tool and library.** Disponível em: <https://curl.haxx.se/>. Acesso em: 04 jul. 2016.

[2] DEVMEDIA. **GitHub e Git - Colaboração e Organização.** Disponível em: <http://www.devmedia.com.br/github-e-git-colaboracao-e-organizacao/24150>. Acesso em: 06 jul. 2016.

[3] GITHUB. **GitHub como as pessoas constroem software.** Disponível em: <https://github.com/>. Acesso em: 03 jul. 2016.

[4] GHTORRENT. **The GHTorrent project.** Disponível em: <http://ghtorrent.org/>. Acesso em: 04 jul. 2016.

[5] GOUSIOS, G. The GHTorent dataset and tool suíte. In: **Proceedings of the 10th Working Conference on Mining Software Repositories**, pp. 233-236. IEEE Press, 2013.

[6] IBM. **Mineração de dados com WEKA, Parte 1: Introdução e regressão.** Disponível em: <http://www.ibm.com/developerworks/br/opensource/library/os-weka1/>. Acesso em: 02 jul. 2016.

[7] SICONFI. **Fundamentos de XBRL.** Disponível em: <https://siconfi.tesouro.gov.br/siconfi/pages/public/conteudo/conteudo.jsf?id=601>. Acesso em: 30 jun. 2016.

[8] SILVA, Paulo Caetano da. **XBRL Framework.** Disponível em: <https://xbrlframework.com/>. Acesso em: 29 jun. 2016.

[9] TABLELESS. **Tudo que você queria saber sobre Git e GitHub, mas tinha vergonha de perguntar.** Disponível em: <http://tableless.com.br/tudo-que-voce-queria-saber-sobre-git-e-github-mas-tinha-vergonha-de-perguntar/>. Acesso em: 07 jul. 2016.

[10] W3C. **Extensible Markup Language (XML) 1.0.** Disponível em: <https://www.w3.org/TR/1998/REC-xml-19980210>. Acesso em: 07 jun. 2016.

[11] WIKIPÉDIA. **A enciclopédia livre.** Disponível em: <https://pt.wikipedia.org/wiki/GitHub>. Acesso em: 04 jul. 2016.

[12] XBRL. **The Business Report Standard.** Disponível em: <https://www.xbrl.org/>. Acesso em: 29 jun