

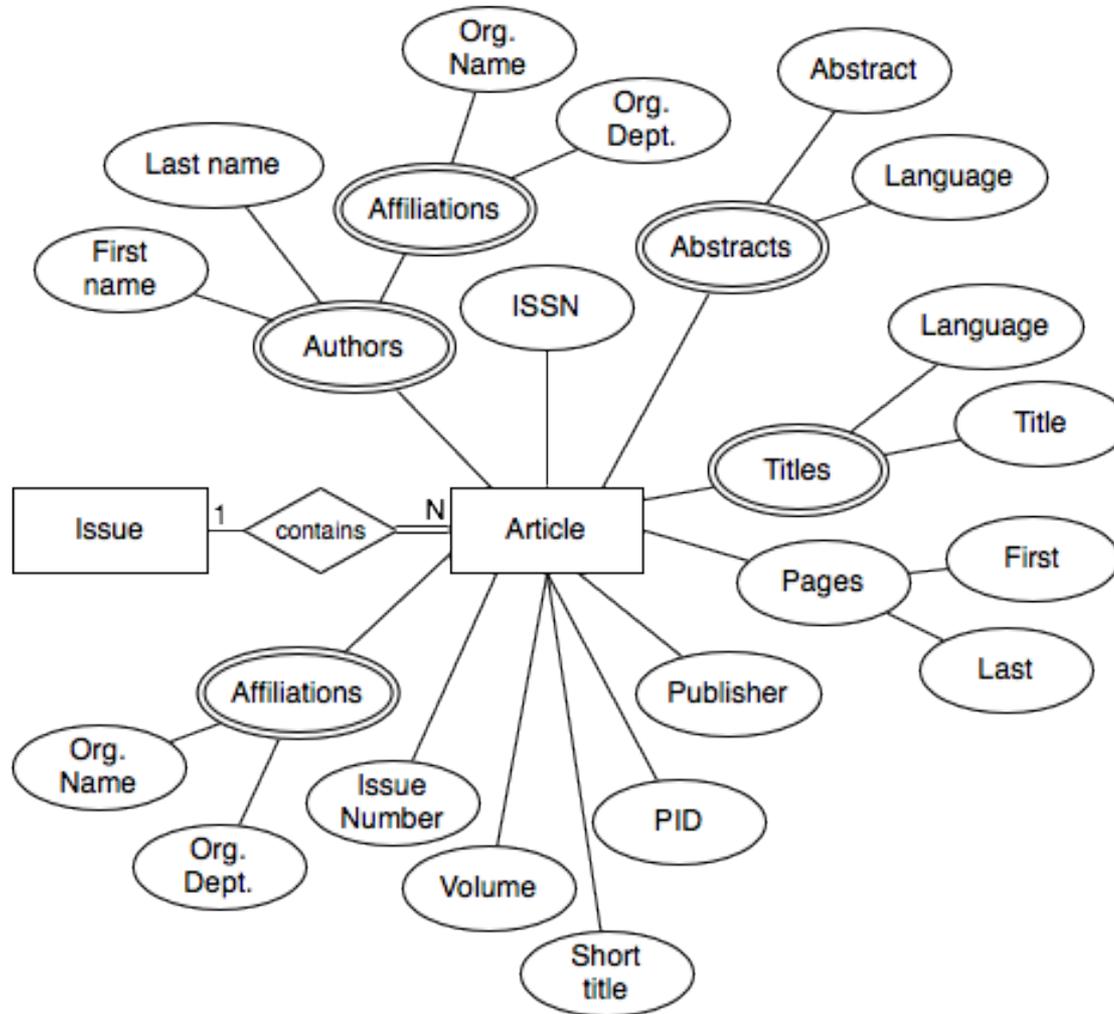
NoSQL

**Panorama atual através das experiências obtidas
no projeto SciELO.**

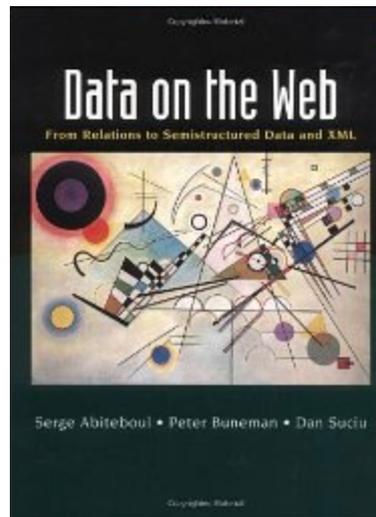
III Encuentro de Desarrolladores de las Redes BVS y Asociadas
por Gustavo Fonseca - 13/09/2012

{fabio.batalha,gustavo.fonseca,jamil.atta}@scielo.org

O modelo semi-estruturado

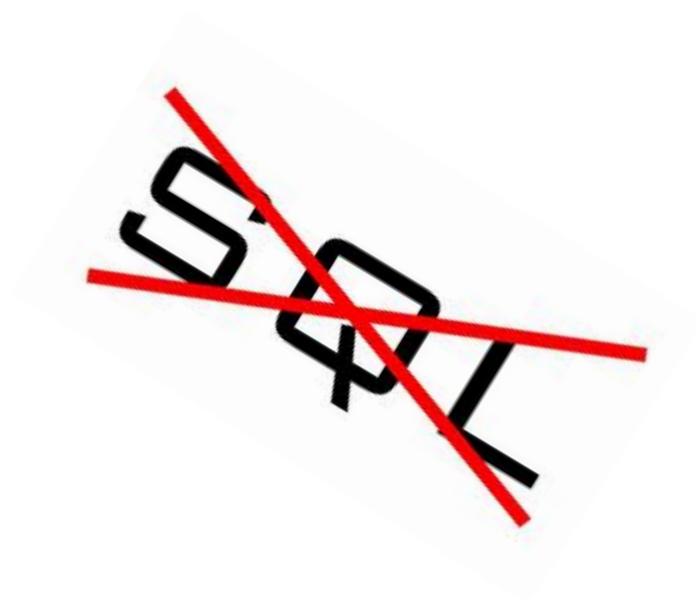


O modelo semi-estruturado



<http://www.amazon.com/Data-Web-Relations-Semistructured-Management/dp/155860622X>

O real significado de NoSQL



Eu uso PostgreSQL, sou um pecador???

Sabores de NoSQL

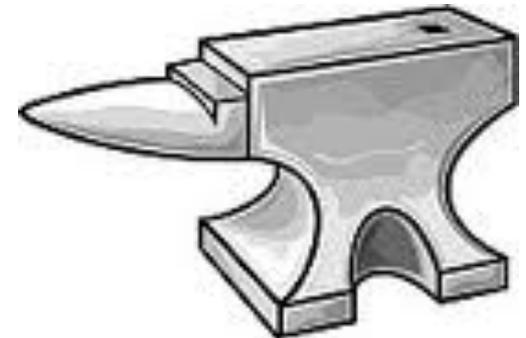
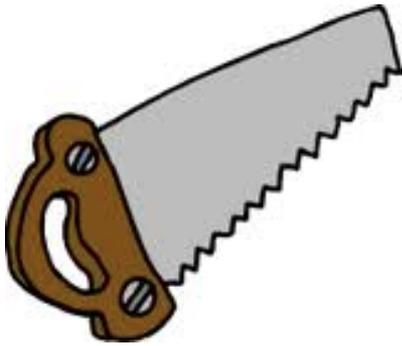
- Documental, Relacional, Grafos, Chave/Valor, Objetos, outros);
- Deve-se considerar a natureza dos dados manipulados;
- Deve-se considerar os requisitos não funcionais da aplicação;
- Provavelmente apenas 1 Sistema de Banco de Dados não resolverá seu problema.

Sabores de NoSQL (<http://nosql-database.org/>)

Wide Column	Document	Key/Value	Graph	Object	XML	Others*
Hadoop	MongoDB	DynamoDB	Neo4J	db4o	Mark Logic	IBM Lotus/Domino
Cassandra	CouchDB	Azure	Infinite Graph	Versant	EMC Documentum xDB	eXtremeDB
Hpertable	RavenDB	Couchbase	Sones	Objectivity	eXist	RDM Embedded
Accumulo	Clusterpoint	Riak	InfoGrid	Starcounter	Sedna	ISIS Family
Amanzon Simple DB	ThruDB	Redis	HyperGraph	Perst	BaseX	Prevalyler
Cloudata	Terrastore	LevelDB	DEX	ZODB	Qizx	Yserial
Cloudera	RaptorDB	Chordless	GraphBase	Magma	Berkeley DB XML	U2
SciDB	JasDB	GenieDB	Trinity	NEO		Globals
HPCC	SisoDB	Scalaris	AllegroGraph	PicoLisp		GT.M
Sytratosphere	SDB	BerkeleyDB	Bigdata	siaqodb		OpenInsight
	SchemaFree	Voldemort	Memorynomy	Sterling		Reality
	djondb	Dynomite	OpenLink	Morantex		OpenQM
		Memcached	VertexDB	EyeDB		ESENT

* Multivalued, Multidimensional, Grid & Cloud, Multimodel, other NoSQL related databases

Qual a melhor solução?



OPS!!!



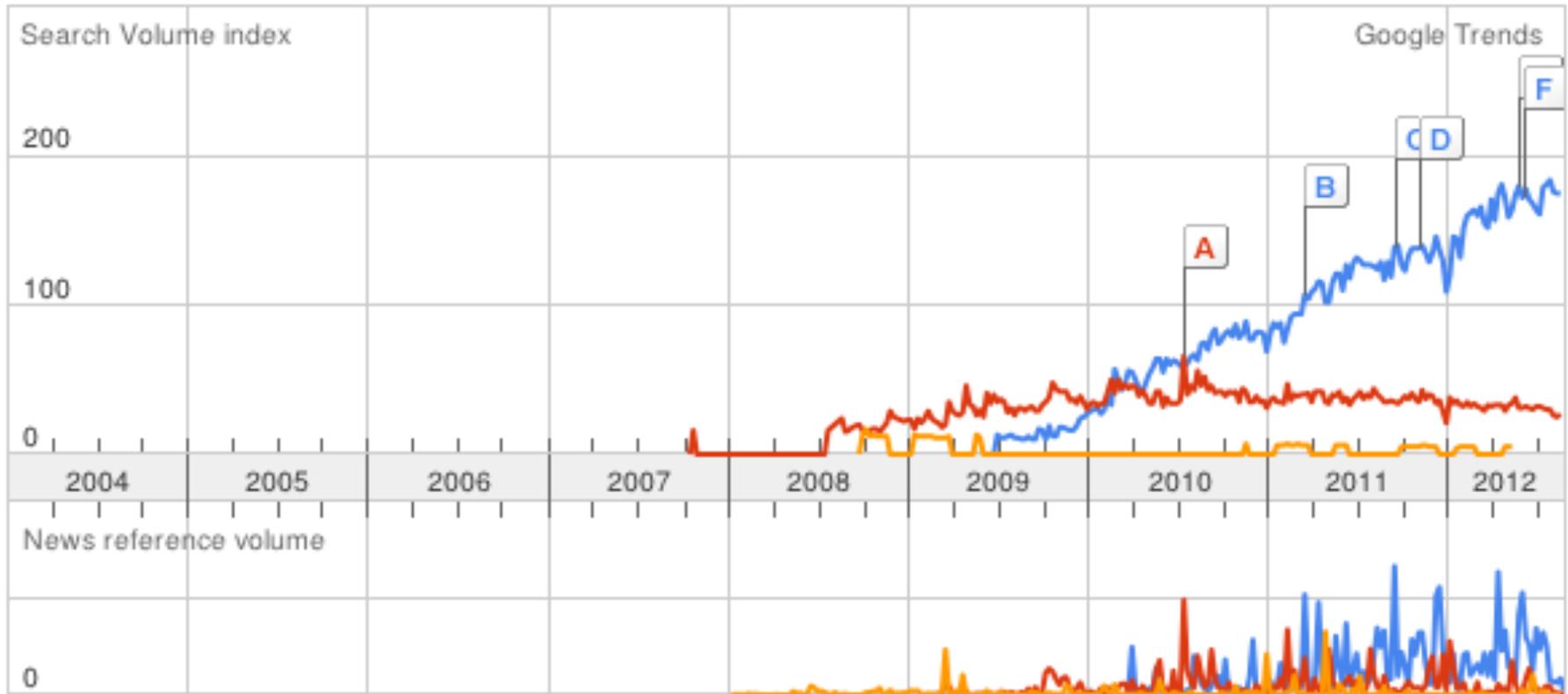
MongoDB

- Criado pela 10gen em 2007
- Primeira versão estável disponível em 2010
- Atualmente um dos bancos de dados NoSQL mais adotados no mercado
- Baixa curva de aprendizado
- Grande comunidade de desenvolvedores e usuários
- API's disponíveis para diversas linguagens de programação (Python, C, C++, Erlang, Haskell, Java, JavaScript, .NET (C#, F#, PowerShell, etc), Node.js, Perl, PHP, Ruby e Scala)

CouchDB

- Criado em 2005 por Damien Katz (ex. Desenvolvedor do Lotus Notes da IBM)
- Em 2008 se tornou um projeto da Apache Software Foundation
- Grande projeção nos últimos 3 anos
- Replicação de dados
- Armazenamento em formato JSON
- Capacidade de desenvolver produtos auto-contidos utilizando CouchApps.
- API's disponíveis para diversas linguagens de programação: C, C#, ColdFusion, Erlang, ExJS, Haskell, Java, JavaScript, LISP, LotusScript, Lua, Node.js, Objective-C, Objective Caml (OCaml), Perl, PHP, PLSQL, Python, Rebol, Ruby e Smalltalk

Bancos de dados documental



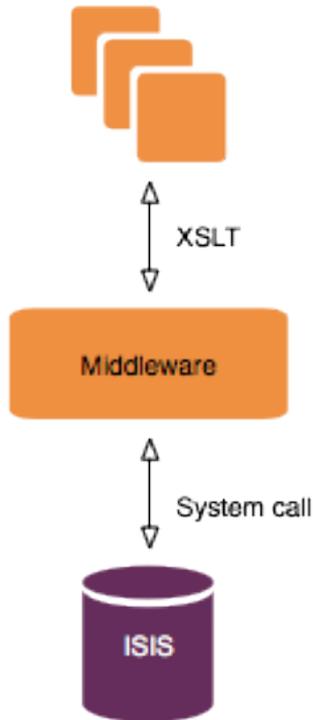
Volume de índices gerados pelo Google para os termos MongoDB (azul), CouchDB (vermelho) e CISIS (amarelo)

Bancos de dados documental

	MongoDB	CouchDB
Escalabilidade horizontal	X	X
Escalabilidade vertical	X	X
Replicação	X	X
Sharding	X	
Tolerância a falhas	X	X
Map-reduce / agregação	X	X
Ordenação	X	X
Pesquisas adhoc	X	
Indexação incremental	X	X
Caracteres Unicode	X	X

	MongoDB	CouchDB
Índice geográfico	X	
Modificadores atômicos	X	
Balanceamento de carga	X	
File storage (BLOBS)	X	X
Indexação	X	X
Integrity	X	X
Concurrency	X	X
Foundation data types	X	X
TTL (time to leave) documents	X	

NoSQL e SciELO



- Basicamente publicação de documentos;
 - artigos, livros;
- Expressividade no modelo de dados;
 - dados semi-estruturados, CISIS;
- Desempenho

Motivações SciELO

- A SciELO gosta de tecnologia
- Open Source
 - Adoção de tecnologias *leading-edge*
 - Aumentar capacidade de contribuição de terceiros
 - Uso de tecnologias que possuem comunidade ativa
 - Documentações atualizadas e existência de vários estudos de casos que servem de referência para desenvolvimentos novos e atuais.
- Escalabilidade
 - A SciELO demanda cada vez mais tecnologias escaláveis
 - Escalabilidade deve ser horizontal e vertical
 - Arquitetura não deve ser um impedimento para escalabilidade

Webservices

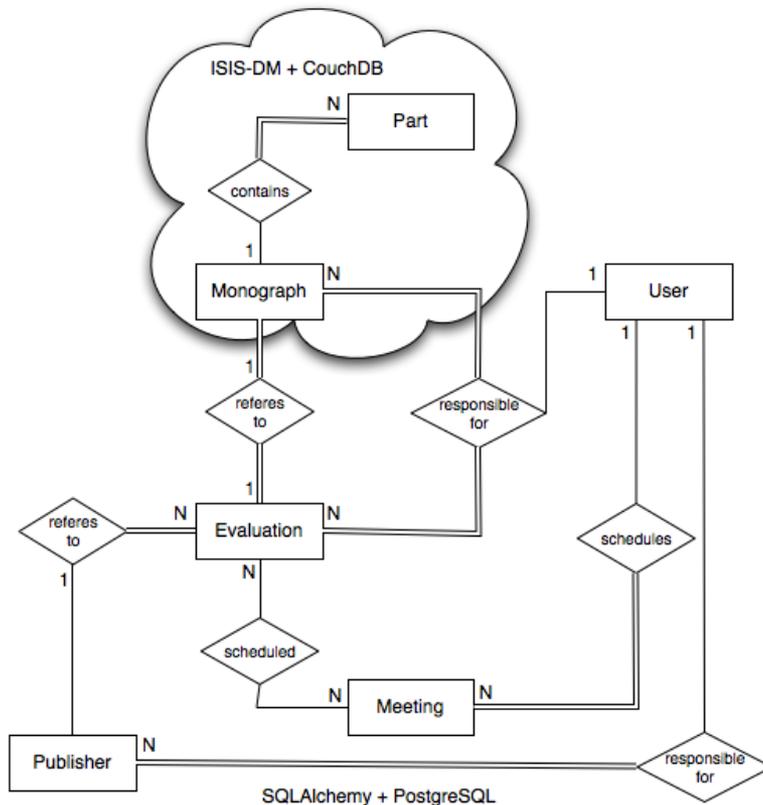
- O precursor na virada tecnológica dos produtos SciELO;
- Prova de conceito do uso do CouchDB com o modelo de dados SciELO;
 - Migração por meio do utilitário isis2json*;
- Utiliza as funcionalidades do CouchDB para expor os dados;
 - CouchApps, Show e View functions.
- Abriu a mente da equipe de desenvolvimento sobre tecnologias de persistência de dados;
- Representou um marco na adoção gradual de uma plataforma baseada em serviços.

* isis2json é um utilitário desenvolvido pela Bireme, para a conversão de bases ISIS para o formato JSON.

Código fonte: <https://github.com/bireme/isis2json>

SciELO Livros

- Catálogo on-line de livros acadêmicos.
- Php, Python, CouchDB, PostgreSQL e MySQL.



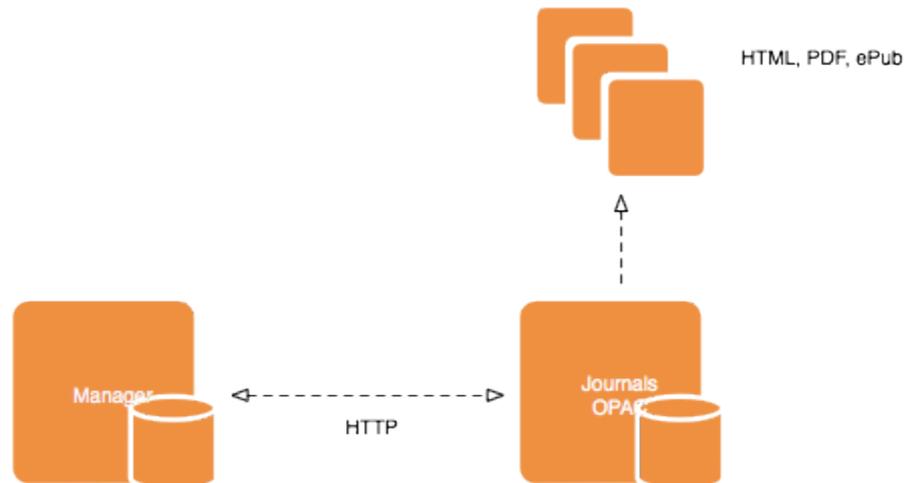
Docs: <http://docs.scielo.org/projects/scielo-books/en/latest/>

Código fonte: <https://github.com/scieloorg/scielobooks>

Licença: BSD-2clause

SciELO Periódicos

- Catálogo on-line de periódicos científicos;
- Atualmente baseado na plataforma ISIS;
- Arquitetura "shared nothing";
- Migração para Python e MongoDB.



Visão geral da nova arquitetura do SciELO Periódicos

Ratchet

- Contador de acessos aos recursos SciELO e associados.
- Python, MongoDB e Tornado.

Leitura*

Nível de concorrência: **250**
Total de requisições: **2000**
Reqs: **1180/sec (média)**
Tempo por req. : **211.033 ms**

Gravação*

Nível de concorrência: **250**
Total de requisições: **2000**
Reqs: **1775/sec (média)**
Tempo por req. : **140.833 ms**

Docs: <http://docs.scielo.org/projects/ratchet/en/latest/>
Código fonte: <https://github.com/scieloorg/ratchet>
Licença: BSD-2clause

* Valores aproximados com base num ambiente GNU/Linux, 8 cores e 32GB RAM. Não foram estabelecidas metodologias sérias para a mensuração, então não leve este benchmark muito a sério. =)

Nurl

- Encurtador de URLs da rede SciELO.
- Python, MongoDB, Memcached, Gunicorn e Nginx.

Resolução*

Nível de concorrência: **100**
Total de requisições: **1000**
Reqs: **219,98 /sec (média)**
Tempo por req. : **454,595 ms**

Encurtamento*

Nível de concorrência: **400**
Total de requisições: **2000**
Reqs: **228,90 /sec (média)**
Tempo por req. : **1747,498 ms**

Docs: <http://docs.scielo.org/en/latest/api/ref-scielo.html>
Código fonte: <https://github.com/scieloorg/nurl>
Licença: BSD-2clause

* Valores aproximados com base num ambiente GNU/Linux, 8 cores e 32GB RAM. Não foram estabelecidas metodologias sérias para a mensuração, então não leve este benchmark muito a sério. =)

Conclusões

- O ecossistema é importante;
- Desempenho é importante, nos 2 contextos:
 - Desenvolvimento (expressividade do programador)
 - Execução da aplicação (ambiente de produção)
- Não existe bala de prata;

Dúvidas?

Obrigado!!

<http://www.scielo.org>

<http://docs.scielo.org>

<http://github.com/scieloorg>

Imagens

Trapped Babe

http://3.bp.blogspot.com/_aoBxVsSWP20/TS2fnNXXuTI/AAAAAAAAAHk/bGwPn_gh684/s1600/busyparents.jpg

Chave de Grifo

https://encrypted-tbn3.google.com/images?q=tbn:ANd9GcQT5eg92le4ZwxwYB6DynhAKLBOQ56iqc40TzK3M4ZGqwew_iGPEw

Martelo

<https://encrypted-tbn0.google.com/images?q=tbn:ANd9GcQS5YLYhwhBwx7VQC6M5LfJX70q9awJ0SM8X9KNqXaM0-34DwfP>

Serra

https://encrypted-tbn0.google.com/images?q=tbn:ANd9GcRTnsfi-lm9Di3JT1vKQa_kJ1xeoP8ehMLic-cT10usy3ais0kNyg

Bigorna

https://encrypted-tbn0.google.com/images?q=tbn:ANd9GcR7GsXWDO47i9kqpB9RqzBiPBG9mFpYs8-l-q464z_1JigbkP4A

Chiclete

<https://encrypted-tbn2.google.com/images?q=tbn:ANd9GcQlpo93lhWcGyNcTjJPhblvv0WhVsFYXGr5u5WHmX7CN6M5LH-0>

Pá

https://encrypted-tbn1.google.com/images?q=tbn:ANd9GcTH5Sh_GA6hqshwZQKZ6RyDiG48DDyQVHmYgS74ivVOoR_tS-ZEmw