# LinkGate

Web Archive Graph Visualization

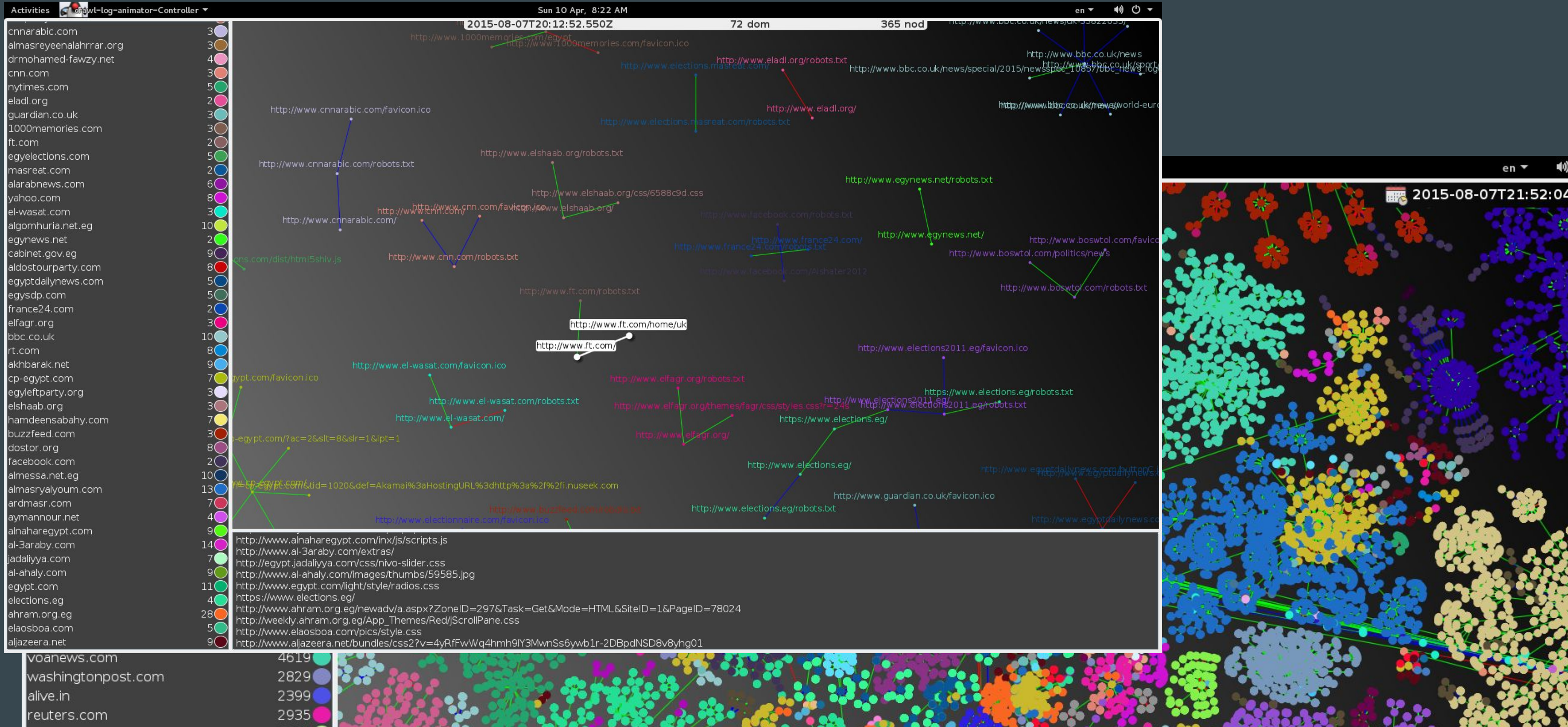
# Rationale, partnership, and IIPC funding

- Visualization is an essential tool for understanding data and conducting research
  - Hyperlinks are the key concept behind the web, making the web a big graph
  - Tools exist for graph visualization, e.g., Gephi, this project addresses scalability and interoperability
- Bibliotheca Alexandrina and National Library of New Zealand worked together to develop core functionality and compile inventory of research use cases
- Funded by the IIPC during 2020

IIPC
netpreserve.org

INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

# Crawl Log Animator



Activities
crawl-log-animator-Controller ▾
Sun 10 Apr, 8:22 AM
en ▾
2015-08-07T20:12:52.550Z
72 dom
365 nod
http://www.1000memories.com/egypt
http://www.1000memories.com/favicon.ico
http://www.elections.masreat.com/
http://www.eladl.org/robots.txt
http://www.bbc.co.uk/news
http://www.bbc.co.uk/sport
http://www.bbc.co.uk/news/special/2015/newsspec_10857/bbc_news_log
http://www.cnnarabic.com/favicon.ico
http://www.eladl.org/
http://www.bbc.co.uk/news/world-eur
http://www.elections.masreat.com/robots.txt
http://www.elshaab.org/robots.txt
http://www.cnnarabic.com/robots.txt
http://www.egynews.net/robots.txt
http://www.elshaab.org/css/6588c9d.css
http://www.cnn.com/favicon.ico
http://www.cnn.com/
http://www.elshaab.org/
http://www.facebook.com/robots.txt
http://www.cnnarabic.com/
http://www.france24.com/
http://www.egynews.net/
http://www.boswtol.com/favico
http://www.france24.com/robots.txt
http://www.boswtol.com/politics/news
ns.com/dist/html5shiv.js
http://www.cnn.com/robots.txt
http://www.facebook.com/Alshater2012
http://www.boswtol.com/robots.txt
http://www.ft.com/robots.txt
http://www.ft.com/home/uk
http://www.ft.com/
http://www.elections2011.eg/favicon.ico
http://www.el-wasat.com/favicon.ico
ypt.com/favicon.ico
http://www.elfagr.org/robots.txt
https://www.elections.eg/robots.txt
http://www.el-wasat.com/robots.txt
http://www.elections2011.eg/
http://www.elections2011.eg/robots.txt
http://www.elfagr.org/themes/fagr/css/styles.css?r=245
http://www.el-wasat.com/
https://www.elections.eg/
http://www.elfagr.org/
-egypt.com/?ac=2&slt=8&slr=1&lpt=1
http://www.elections.eg/
http://www.egyptdailynews.com/button
http://www.egyptdailynews
ww.cp-egypt.com/
l=cp-egypt.com&tid=1020&def=Akamai%3aHostingURL%3dhttp%3a%2f%2fi.nuseek.com
http://www.guardian.co.uk/favicon.ico
http://www.elections.eg/robots.txt
http://www.electionnaire.com/favicon.ico
http://www.egyptdailynews.co
en ▾
2015-08-07T21:52:04

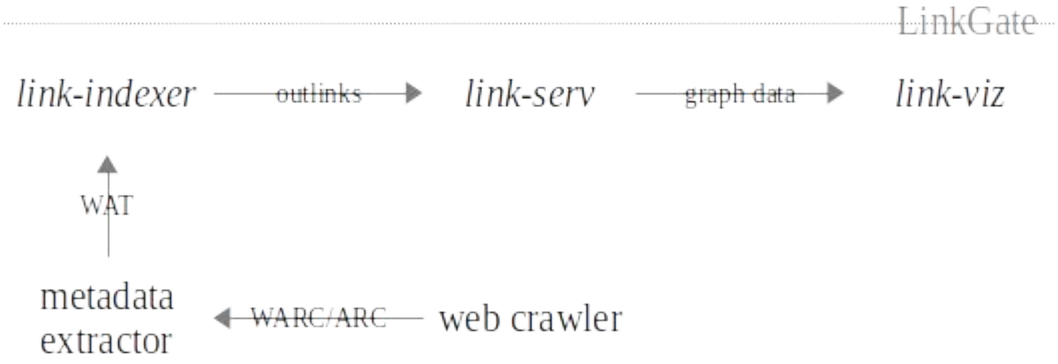
| | |
|---|---|
| cnnarabic.com | 3 |
| almasreyeenalahrrar.org | 3 |
| drmohamed-fawzy.net | 4 |
| cnn.com | 3 |
| nytimes.com | 5 |
| eladl.org | 2 |
| guardian.co.uk | 3 |
| 1000memories.com | 3 |
| ft.com | 2 |
| egyelections.com | 5 |
| masreat.com | 2 |
| alarabnews.com | 6 |
| yahoo.com | 8 |
| el-wasat.com | 3 |
| algomhuria.net.eg | 10 |
| egynews.net | 2 |
| cabinet.gov.eg | 9 |
| aldostourparty.com | 8 |
| egyptdailynews.com | 5 |
| egysdp.com | 5 |
| france24.com | 2 |
| elfagr.org | 3 |
| bbc.co.uk | 10 |
| rt.com | 8 |
| akhbarak.net | 9 |
| cp-egypt.com | 7 |
| egyleftparty.org | 3 |
| elshaab.org | 3 |
| hamdeensabahy.com | 7 |
| buzzfeed.com | 3 |
| dostor.org | 8 |
| facebook.com | 2 |
| almessa.net.eg | 10 |
| almasryalyoum.com | 13 |
| ardmasr.com | 7 |
| aymannour.net | 4 |
| alnaharegypt.com | 9 |
| al-3araby.com | 14 |
| jadaliyya.com | 7 |
| al-ahaly.com | 9 |
| egypt.com | 11 |
| elections.eg | 4 |
| ahram.org.eg | 28 |
| elaosboa.com | 5 |
| aljazeera.net | 9 |

```
http://www.alnaharegypt.com/inx/js/scripts.js
http://www.al-3araby.com/extras/
http://egypt.jadaliyya.com/css/nivo-slider.css
http://www.al-ahaly.com/images/thumbs/59585.jpg
http://www.egypt.com/light/style/radios.css
https://www.elections.eg/
http://www.ahram.org.eg/newadv/a.aspx?ZoneID=297&Task=Get&Mode=HTML&SiteID=1&PageID=78024
http://weekly.ahram.org.eg/App_Themes/Red/jScrollPane.css
http://www.elaosboa.com/pics/style.css
http://www.aljazeera.net/bundles/css2?v=4yRfFwWq4hmh9lY3MwnSs6ywb1r-2DBpdNSD8v8yhg01
```

| | |
|---|---|
| voanews.com | 4619 |
| washingtonpost.com | 2829 |
| alive.in | 2399 |
| reuters.com | 2935 |

# Overview

- 3 components:
  - *link-serv*
  - *link-indexer*
  - *link-viz*
- link-serv: scalable graph data service
- link-indexer: get link data from web archive storage and insert into link-serv
- link-viz: visualize and explore graph data from link-serv inside a web browser
- Inventory of research use cases to guide futur development



# *link-viz*: Web Archive Graph Visualization Frontend

# *link-indexer*: Linked Data Collection Tool

# What is *link-indexer*?

- Data flow begins with *link-indexer*
- From input data (e.g., WAT files), for each record, extract the following at a minimum:
  - Identifier/URI
  - Timestamp
  - Outlinks
- Produces graph data to insert into *link-serv*
- Written in Python
- Uses *webarchive-commons*, *urlcanon*, and *warcio* from the web archiving tool ecosystem

# *link-indexer* features

- Input format handling implemented as separate modules
- Generate WAT on the fly
- Post data to API endpoint or dry-run for troubleshooting
- Improve performance via batch processing
- Configurable network tolerance
- Script-friendly logging output
- Configurable input data handling behavior
- Configurable error handling behavior
- Load options from a configuration file

# What’s next?

- Continue large-scale testing
- Enhance logging, including providing extra details, writing to a remote logging service
- Support more input formats
- More command-line options
- Extract additional metadata

# *link-serv*: Temporal Graph Data Service

# What is *link-serv*?

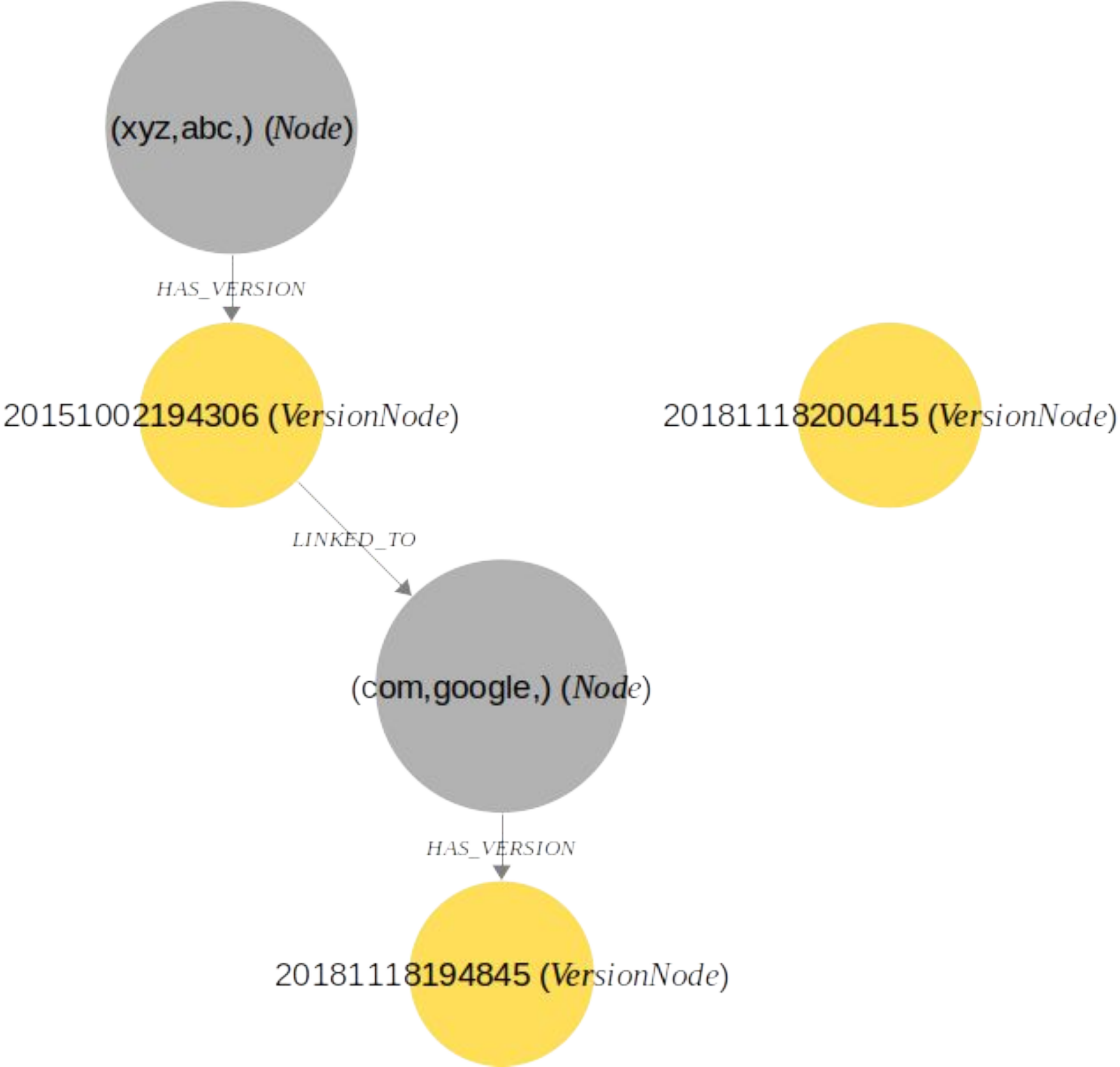
- A service to provide a RESTful API for inserting temporal graph data extracted from a web archive into a central data store
- This service is used for retrieving back that data for rendering and navigation
- Design goals:
  - Data store scalability
  - Use publicly licensed technology
  - Data schema for temporal (i.e., versioned) graph data
  - RESTful API and Gephi compatibility

# Graph data stores

- A technical survey for data stores has been conducted
- Virtuoso, 4Store, OrientDB, Neo4j, ArangoDB, and others were considered
- Neo4j and ArangoDB are chosen as the best fit for the application
- Both provide high performant read and write operations

# Neo4j data model

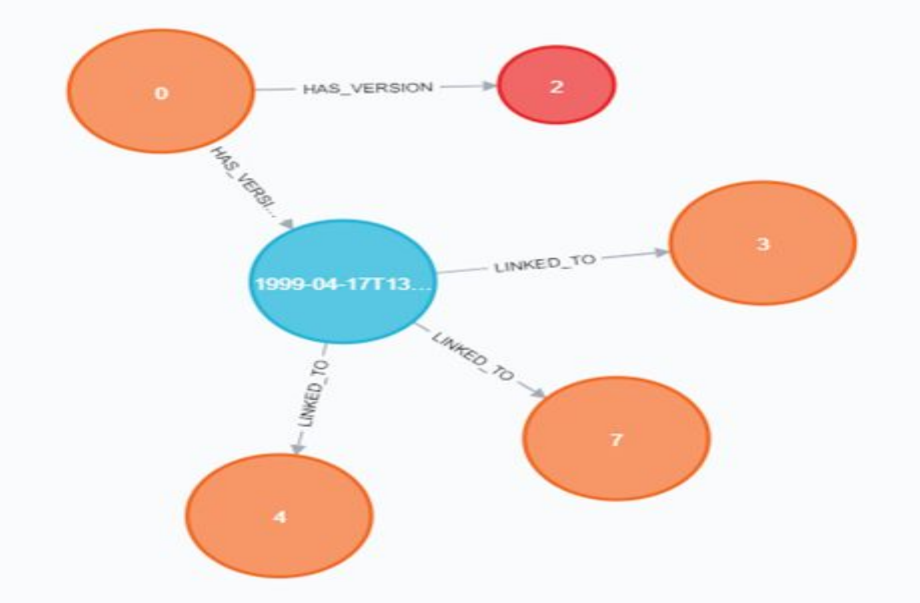
-


(xyz,abc,) (Node)
HAS_VERSION
20151002194306 (VersionNode)
20181118200415 (VersionNode)
LINKED_TO
(com,google,) (Node)
HAS_VERSION
20181118194845 (VersionNode)

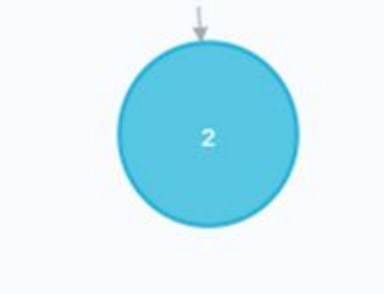
# Neo4j data model

- This mainly depends on the concept of “edgecuts”

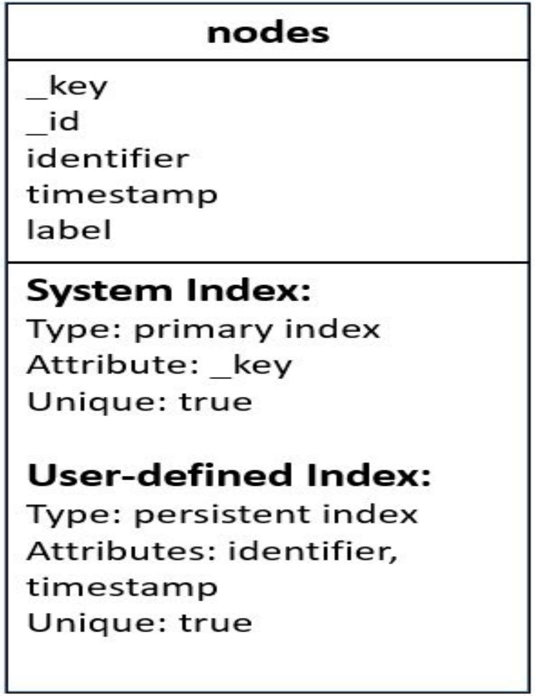
Shard 1



Shard 2



# ArangoDB data model

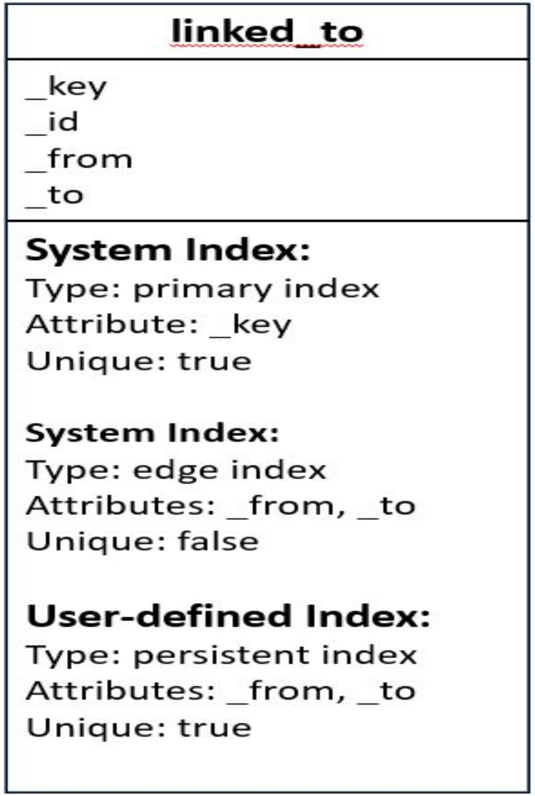


**nodes**

_key
_id
identifier
timestamp
label

**System Index:**
Type: primary index
Attribute: _key
Unique: true

**User-defined Index:**
Type: persistent index
Attributes: identifier, timestamp
Unique: true



**linked_to**

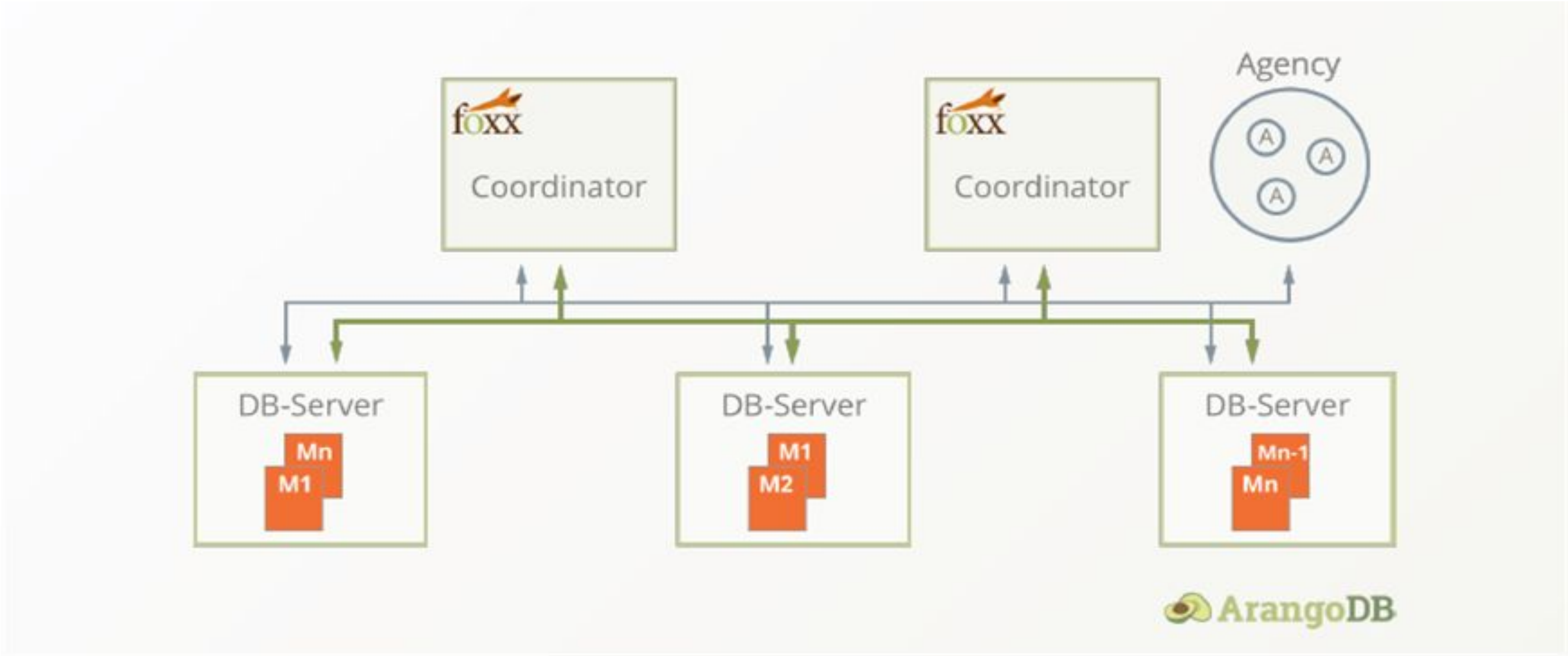
_key
_id
_from
_to

**System Index:**
Type: primary index
Attribute: _key
Unique: true

**System Index:**
Type: edge index
Attributes: _from, _to
Unique: false

**User-defined Index:**
Type: persistent index
Attributes: _from, _to
Unique: true

# ArangoDB clustering

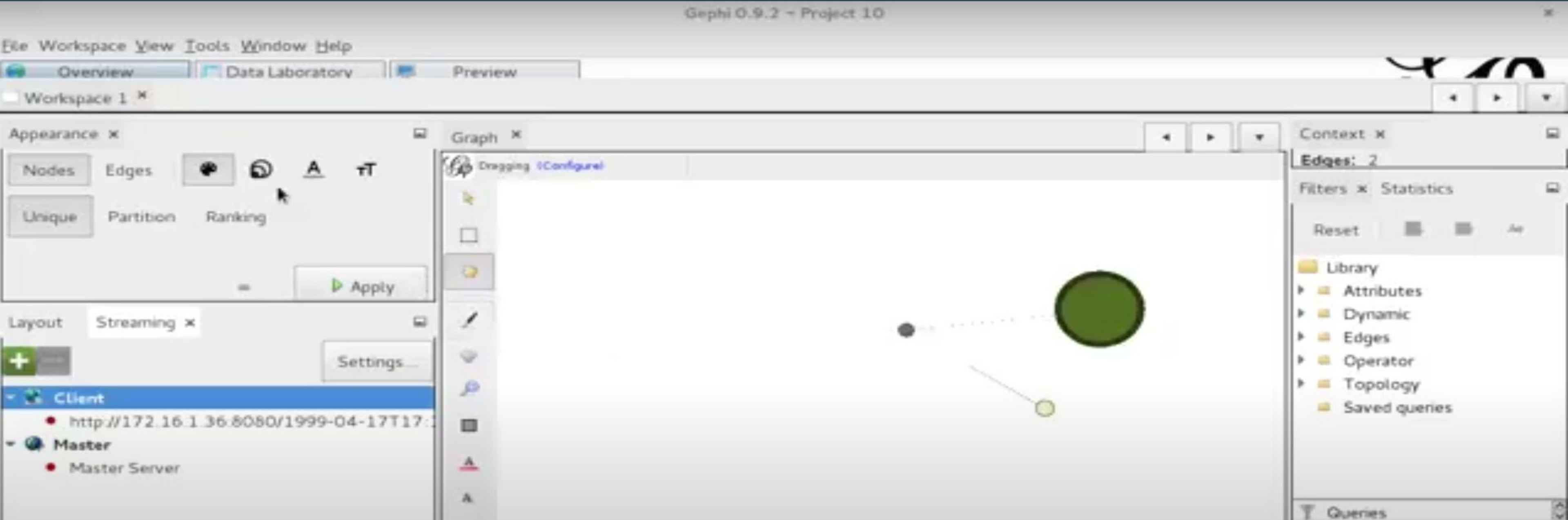


# API

- *link-serv* is implemented as a web service using Java and the Spring framework
- API for exposing functionality provided by the data model
- Currently implemented operations:
  - updateGraph
  - getGraph
  - getVersionCountsYearly
  - getVersionCountsMonthly
  - getVersionCountsDaily
  - getVersions
  - getLatestVersion

# Gephi compatibility

- *link-serv* is compatible with the API used by the Gephi streaming plugin
- Gephi can be used to render data from *link-serv*



http://172.16.1.36:8080/1999-04-17T17:

# What’s next?

- Design and implementation of replication mechanism according to the expected and real workload
- Large scale testing is a challenge for both data store solutions
- Enhance logging
- Support alternative data stores

# Research Use Cases for Web Archive Graph Visualization

# Use cases to guide future development

- Tracking the promulgation of content through a web archive
- Providing tailored viewshafts into web archives
- Tagging and grouping web archive content with attributes
- Visualizing images and texts
- Preprocessing links before loading into visualization software
- Creating visualizations of crawl log data
- Creating curated web archive views for classrooms, or different audiences

Inventory of use cases: https://github.com/arcalex/linkgate/wiki/Use-cases

LINKGATE

LinkGate:
https://linkgate.bibalex.org

Stay tuned for updates:
https://netpreserveblog.wordpress.com/tag/linkgate/

Get in touch:
linkgate@iipc.simplelists.com

On GitHub:
https://github.com/arcalex/linkgate
https://github.com/arcalex/link-serv
https://github.com/arcalex/link-indexer
https://github.com/arcalex/link-viz