Geo-Crawler

Prepared by

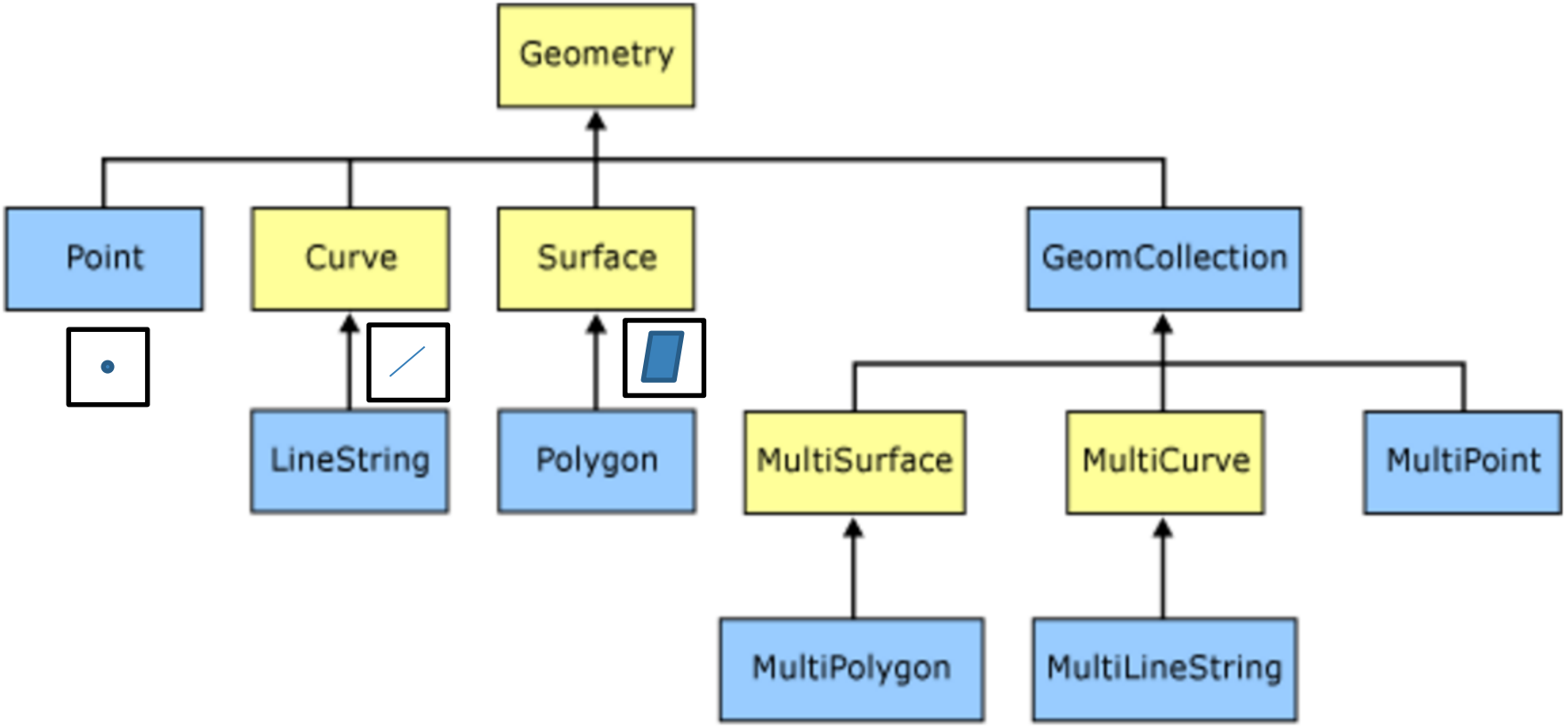
Deepak Punjabi (15IT60R17)

Under Guidance of

Professor Soumya K. Ghosh

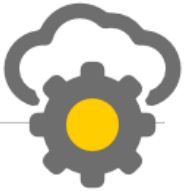
Spatial data
is data containing Information about the locations and shapes of geographic features and the relationships between them, usually stored as coordinates and topology.

Source: http://support.esri.com/



Spatial Object Types

Source: msdn.microsoft.com



OGC Web Services

WMS

- Deliver map images
- Metadata about available layers
- GetCapabilities, GetMap, DescribeLayer

WFS

- Direct access to features
- GML/SOAP interface
- Query/get feature
- Add feature
- Delete feature
- Update feature

WCS

- Multi-dimensional coverage of data
- Provides sptio-temporal information
- Provides rich semantics than WMS and WFS

3 *terabytes*

on daily basis

Source : [NASA:2007]

How to search spatial data ?

Catalog Approach

- Registry not up to date
- Incorrect classification of services
- Not all service providers registers, all kind of services

Utilize popular search engines

- Google, Yahoo, Bing etc.
- Uses page rank, instead of quality of service (QoS)

What is a *Crawler* ?

" A program that systematically browses the
World Wide Web
in order to create an index of data. "

- E.g. bingbot, polybot, googlebot

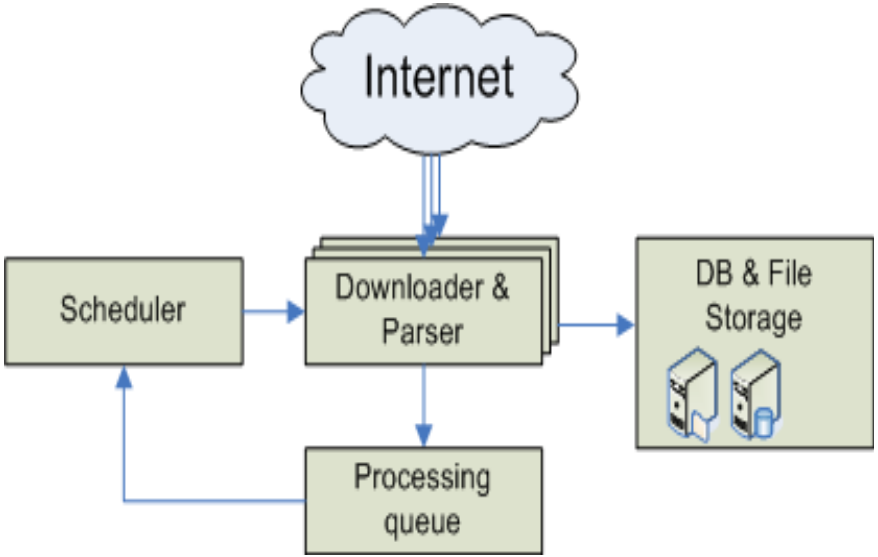
Challenges

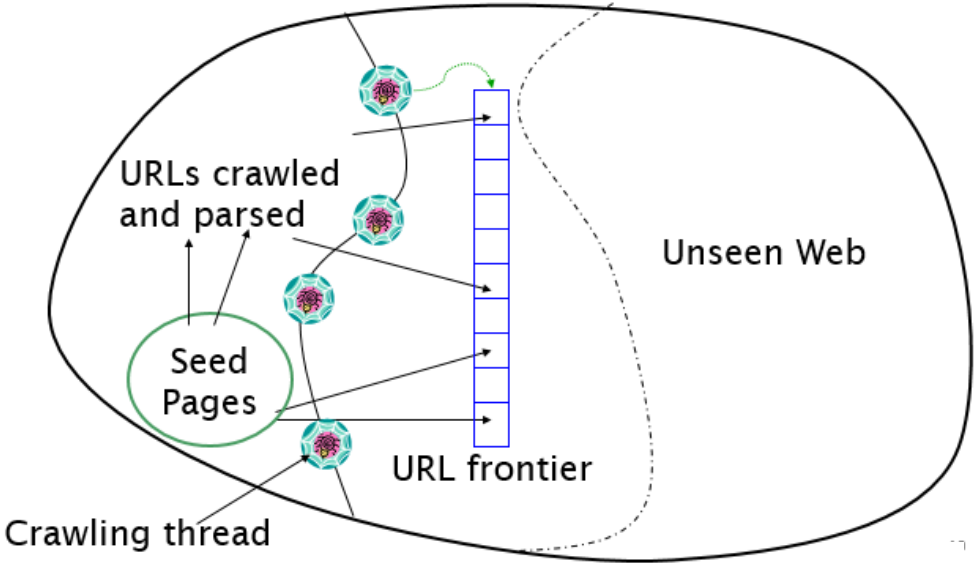
- Scale of the web
- Refresh rate
- heterogeneity

Types of Crawler

- Universal crawler
- Focused crawler
- Topical crawler

How it works ?





Source: nazou.fiit.stuba.sk/home/?page=webcrawler

Spatial Web Crawler: Objectives

- **Building a spatial web crawler** using *WFS* based on *OGC* standard.
- Building a ***domain ontology*** with spatial *feature type*.
- **Semantic matching** using *ontology* and indexing of geo-servers with offered *feature type* reference.
- Performing experiment with test seed *URLs* and **analysing the performance** of the crawler in terms of accurate semantic annotations.

Crawler Architecture

Extraction module

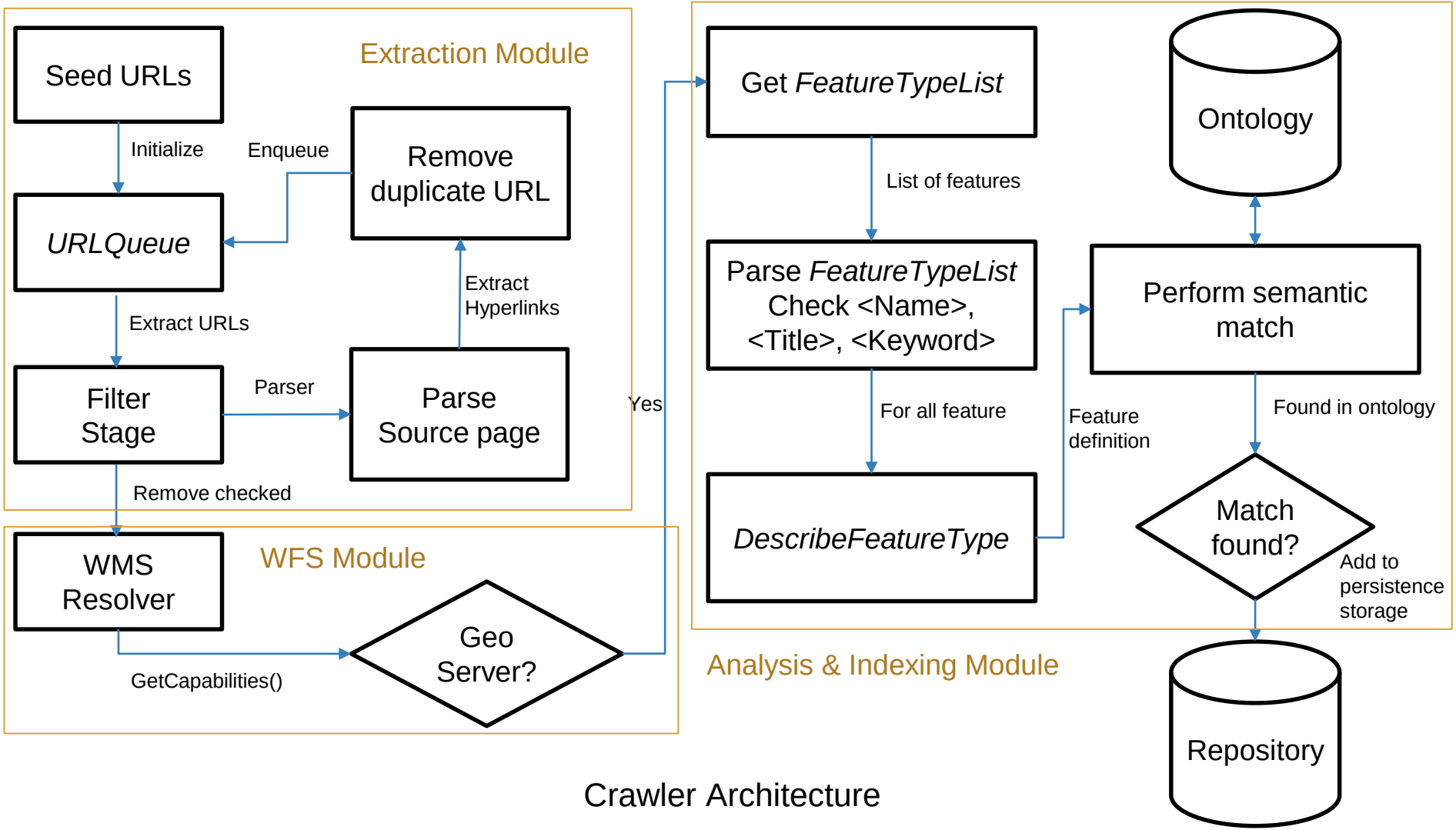
- Read URL from *URLQue*
- Extract hyperlinks
- Remove duplicates
- Push to *URLQueue*

WFS module

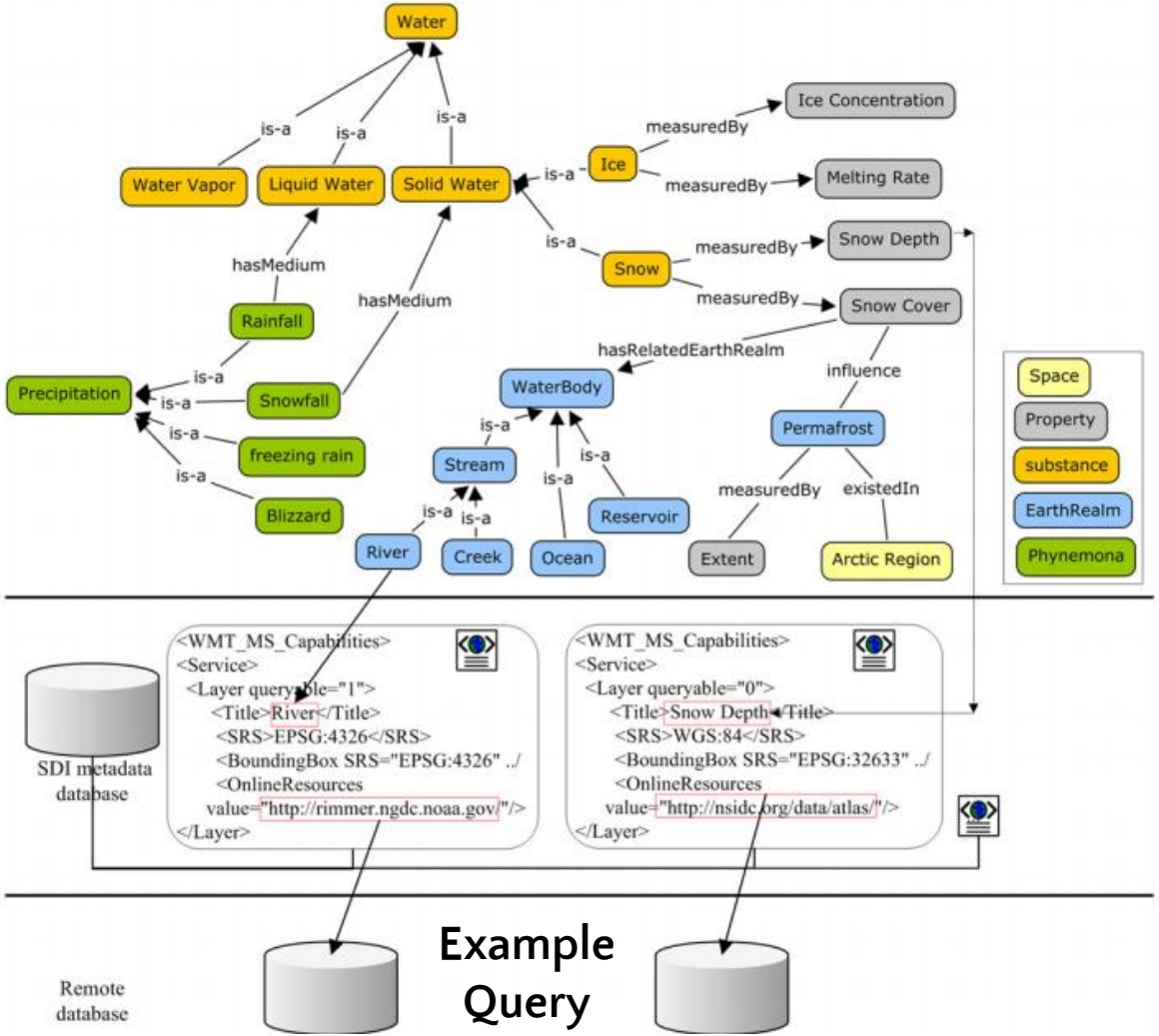
- Generate GetCapabilities request by appending to URL
- Check whether server is a *WFS* server via XML response

Analysis & Indexing module

- Extract *features*
- Perform a semantic match
- Compare extracted features with *ontology*
- Add geo-server to repository



Crawler Architecture



source: Li, W., *et al*. Computers & Geosciences,2011.

Advantages of Spatial web crawler

- Allows searching of pages that are currently not searchable from the general search engines
- Provides a more up-to-date search
- Provides improved accuracy and extra features not possible with general search engines

Performance Evaluation

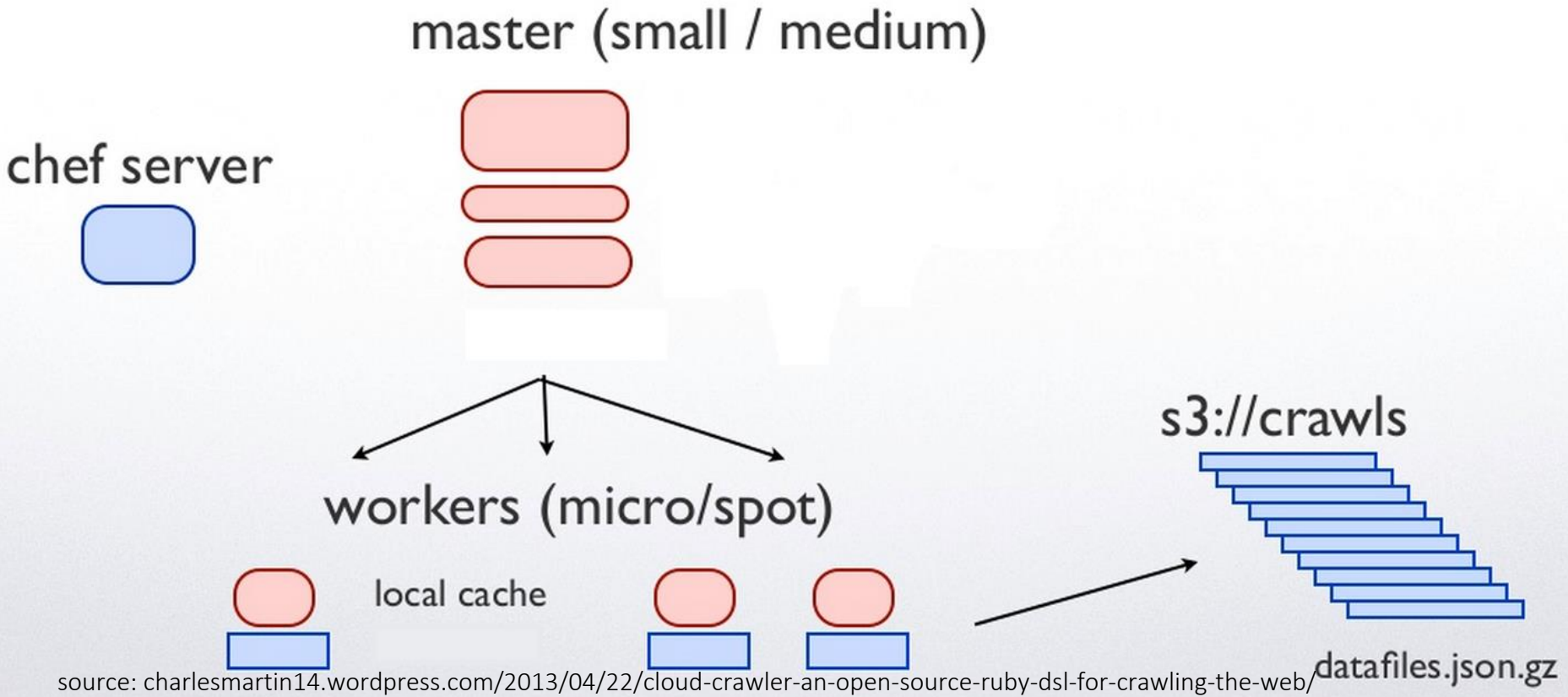
- $$precision = \frac{(Number\_of\_relevant\_geoservers\_found)}{(Total\_Number\_of\_geoservers\_found)} * 100\%$$
- $$recall = \frac{Number\_of\_relevant\_geoservers\_found\_in\_search}{Total\_Number\_of\_existing\_relevant\_geoservers} * 100\%$$
- $$F1 = 2 * \frac{(precision * recall)}{(precision + recall)}$$

Final score is calculated by taking average over all *feature types*.

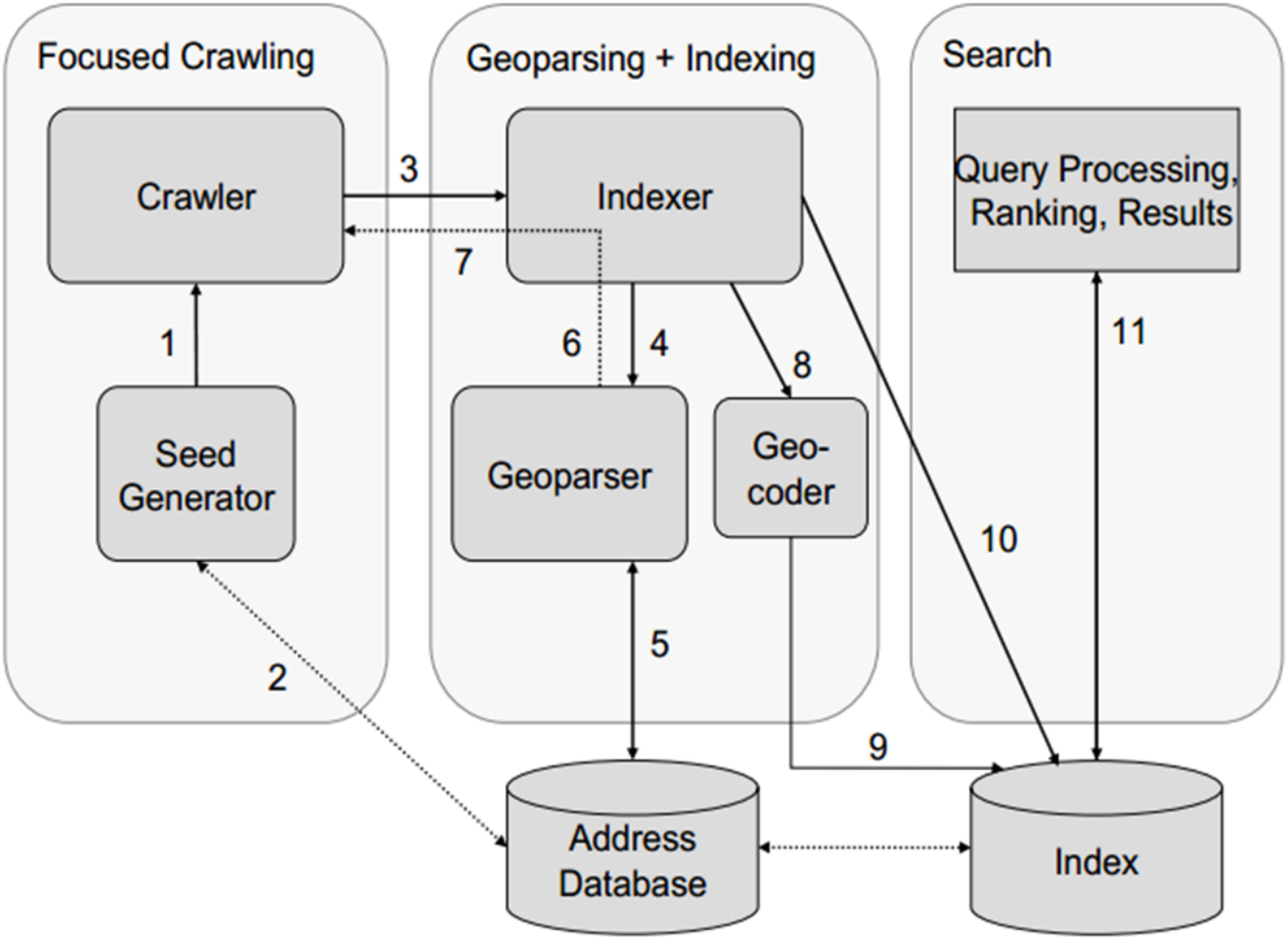
Future work & Extensions

- Priority based crawling
- Parallelization
- Cloud based crawler implementation
- Spatial search engine & ranking

cloud-crawler: architecture



source: charlesmartin14.wordpress.com/2013/04/22/cloud-crawler-an-open-source-ruby-dsl-for-crawling-the-web/



Architecture of a spatial search engine [Source: Ahlers *et al.* Springer London, 2009.]

References

I. Patil, Sonal, Shrutilipi Bhattacharjee, and Soumya K. Ghosh. "**A spatial web crawler for discovering geo-servers and semantic referencing with spatial features.**" Distributed Computing and Internet Technology. Springer International Publishing, 2014. 68-78.

II. Li, Wenwen, Chaowei Yang, and Chongjun Yang. "**An active crawler for discovering geospatial web services and their distribution pattern–a case study of OGC web map service.**" International Journal of Geographical Information Science 24.8 (2010): 1127-1147.

III. Jiang, Jun, Chong-jun Yang, and Ying-chao Ren. "**A spatial information crawler for opengis wfs.**" Sixth International Conference on Advanced Optical Materials and Devices. International Society for Optics and Photonics, 2008.

IV. Marc Najork. "**Web crawler architecture.**" Microsoft Research.

V. Ahlers, Dirk, and Susanne Boll. "**Location-based Web search**." The Geospatial Web. Springer London, 2009. 55-66.

VI. Li, W., et al. "**Semantic-based web service discovery and chaining for building an Arctic spatial data infrastructure.**" Computers & Geosciences 37.11 (2011): 1752-1762.

Thanks!

Any questions ?