Mining Passively Crowdsourced Medical Information

Data|Hack|Award|2014

Draft Proposal 2014-04-12

Danny Ayers

Summary

- discover blogs (and certain social networks) containing posts in the medical domain
- aggregate content
- identify and extract key information
- express the results as linked data using established vocabularies
- display results via Web UI

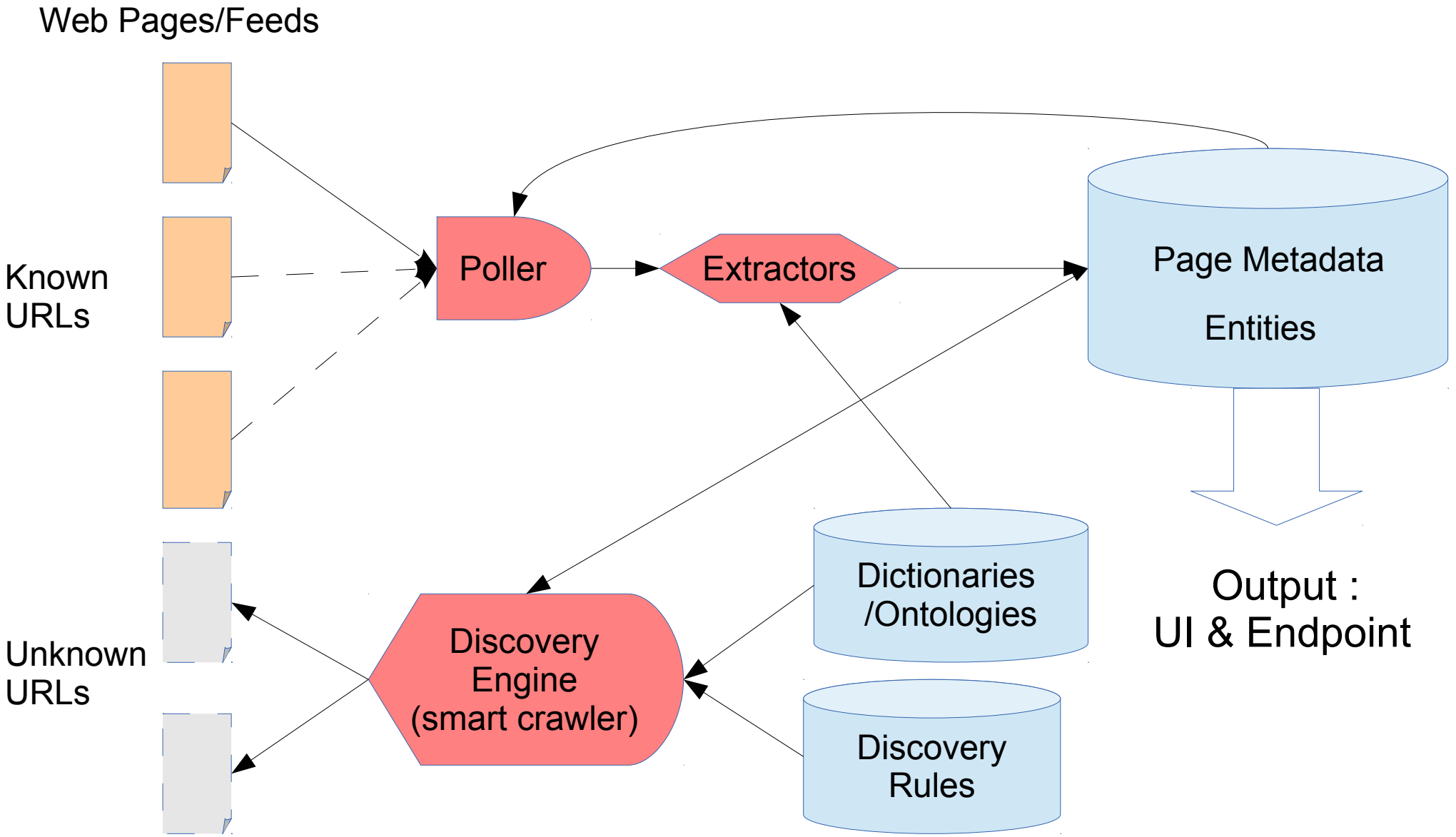
Why blogs/social?

- more human-digestible than typical medical material
- generated in real time, without the latency found in formal research
- not constrained by national borders, open, free

Drawbacks?

- informal language
- less rigorous than conventional medical research
- potential legal/copyright complications

System Overview



Output

- aggregated posts, keyword searchable
- linked data (using established vocabs)
- SPARQL Endpoint
- RESTful interfaces to system components as appropriate

Potential Users

- medical professionals
- pharmaceutical industry
- lay users
- journalists

Use Cases

- general user (patient) reference
- medical news aggregation
- indication of potential drug side effects
- unanticipated reuse...

Rough Roadmap (seriously iterative)

- identify existing Fusepool/Stanbol components that fulfil parts of the requirements
- hack prototypes of missing components, virtually standalone
- minimum viable product... (i.e. get it working!)
- replace/refactor prototypes components into OSGi and integrate with Fusepool
- custom UI (if necessary)
- Live demo deployment

(tests & docs continuous throughout)

Milestones

- Week 1 : identify usable Fusepool/Stanbol components, clarify requirements & design
- Week 2 : prototype poller
- Week 3 : prototype crawler
- Week 4 : prototype system integration
- Week 7 : completed components
- Week 8 : live demo

Future Enhancements

- real-time actions triggered in response to search queries
- user interaction, voting up/down sources
- per-user customization
- UI integration with more of LOD cloud
- enrichment with schema.org, RDFa annotation of human-oriented views

Legacy for Fusepool

- Showcase App
- Application reusable for different domains
- Poller/Aggregator component(s)
- Intelligent Discovery Engine component

Open Questions

- *time, cost, scope..?*
- existing components?
- custom UI?
- discovery rules engine : use RDFS/OWL reasoning? SPARQL? Drools?
- copyright issues?
- *subproject name(s)..?*

Addendum

after discussions 2014-04-12

- change of domain - instead of medical, change to material related to European Parliament
- starting point : MEPs
- initial data/vocabs already available
- should be easier to bootstrap the system as the domain is more constrained
- can also experiment with medical domain to check requirements for redeployment