A Planetary Defense Gateway for Smart Discovery of relevant Information for Decision Support

Myra Bambacus/GSFC & Chaowei Phil Yang/GMU

Ronald Y. Leung, Brent Barbee, Joseph A. Nuth, and Bernie Seery
NASA Goddard Space Flight Center (GSFC)

Yongyao Jiang, Han Qin, Yun Li, Manzhu Yu, Mengchao Xu
George Mason University (GMU)

David S. P. Dearborn
Lawrence Livermore National Laboratory

Catherine Plesko
Los Alamos National Laboratory

- Background
- Framework architecture
- Current results
- Ongoing research
- Conclusions

Planetary Defense (PD)

- Near Earth object (NEO) observation
- Design reference asteroids
- Impact modelling
- Decision support
- Mitigation action

- In this U.S., the NASA Planetary Defense Coordination Office (PDCO) was established in 2016 to study the mitigation of potential Near-Earth Object (NEO) impacts to our home planet.

Image source: http://www.universetoday.com/128347/nasa-discovers-72-new-never-seen-neos/

Motivation for an Information Framework

- Information about detecting, characterizing and mitigating NEO threats is dispersed (e.g. publications, briefings.)

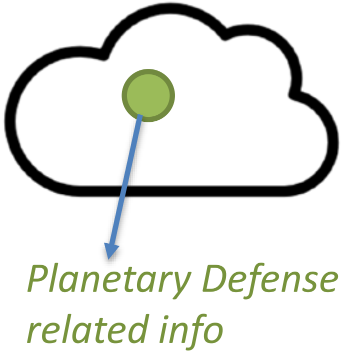
- An overall architecture to facilitate the collaborations and integrate the different capabilities to achieve the most sensible, executable options for mitigation

- A cyberinfrastructure to capture mitigation trades, analyses, model output, risk projections, and mitigation mission design concepts

- Discovery and easy access to knowledge and expert opinion within the project team, as well as factoring in related information from other research and analysis activities

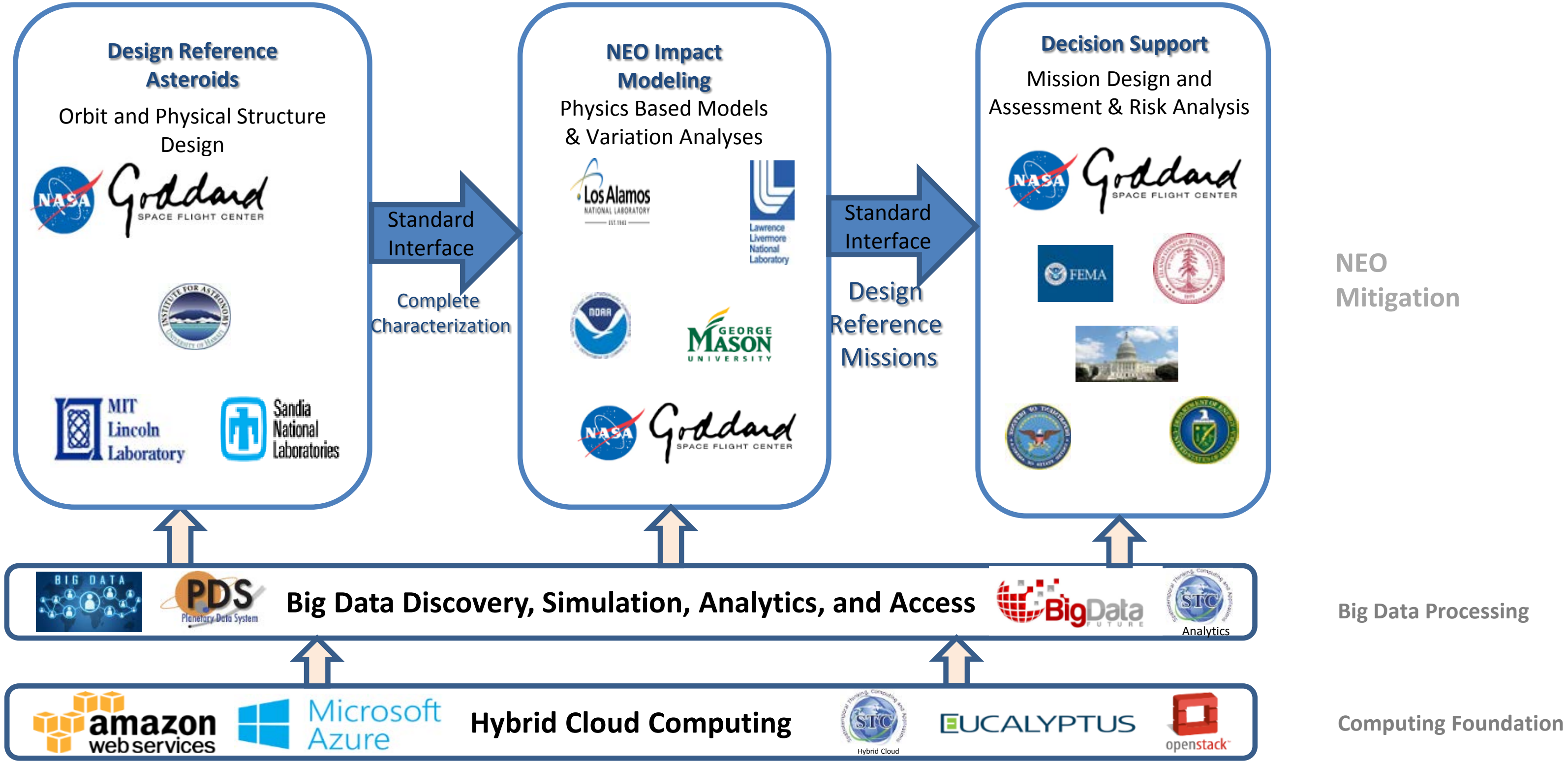
Why Another Resource Discovery Engine?

- Domain-specific vs. general-purpose



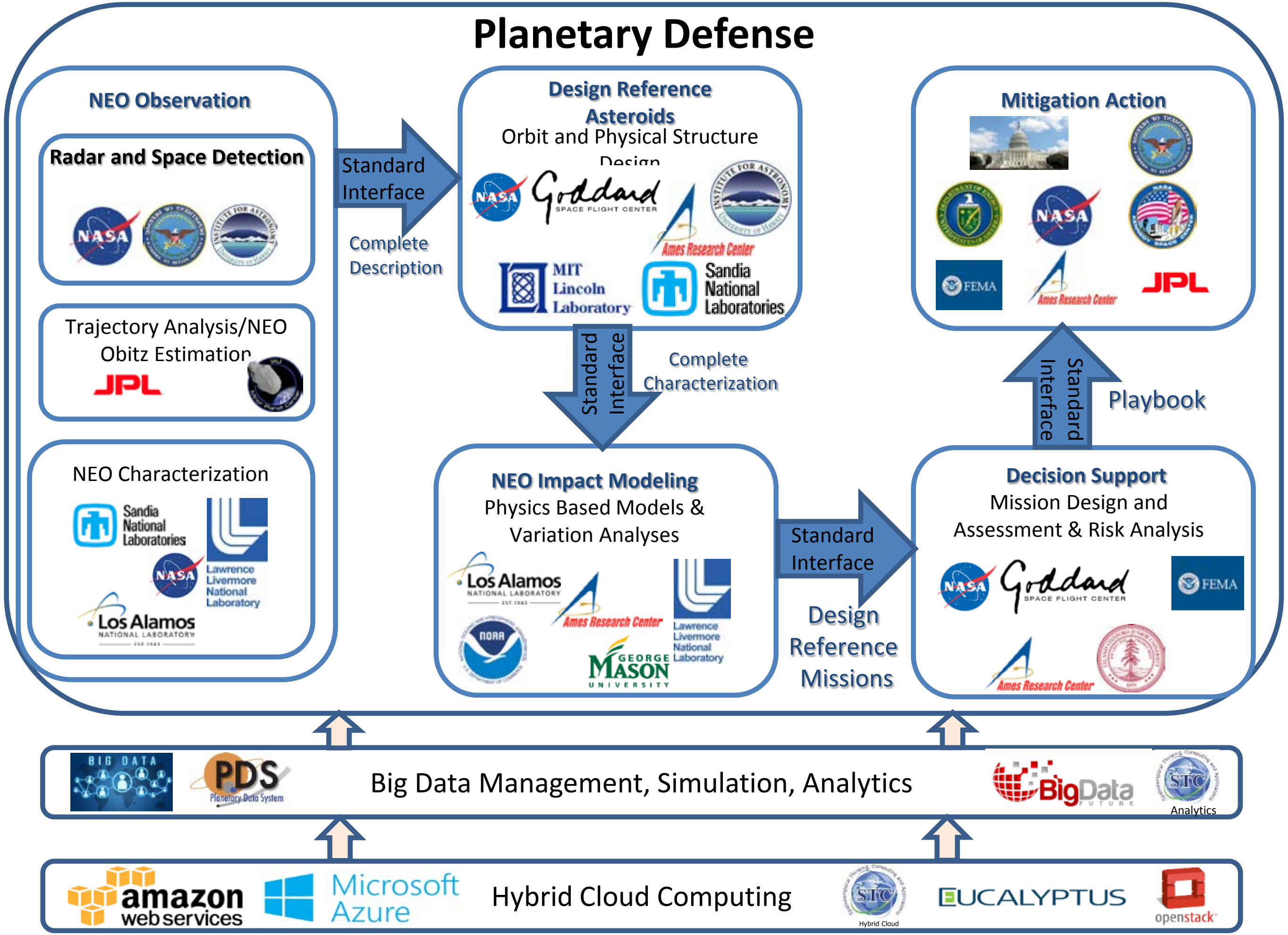
- Indexed content
  - Google searches from nearly the entire Internet
  - The framework is PD-specific
- Knowledge base
  - Google's *Knowledge Graph* is based on generic sources such as Wikipedia
  - The framework will create a PD ontology aided by domain experts, combined with machine learning and Natural Language Processing (NLP) results
- Decision makers can have easy access to required information and quality knowledge

Project Organizational Collaboration

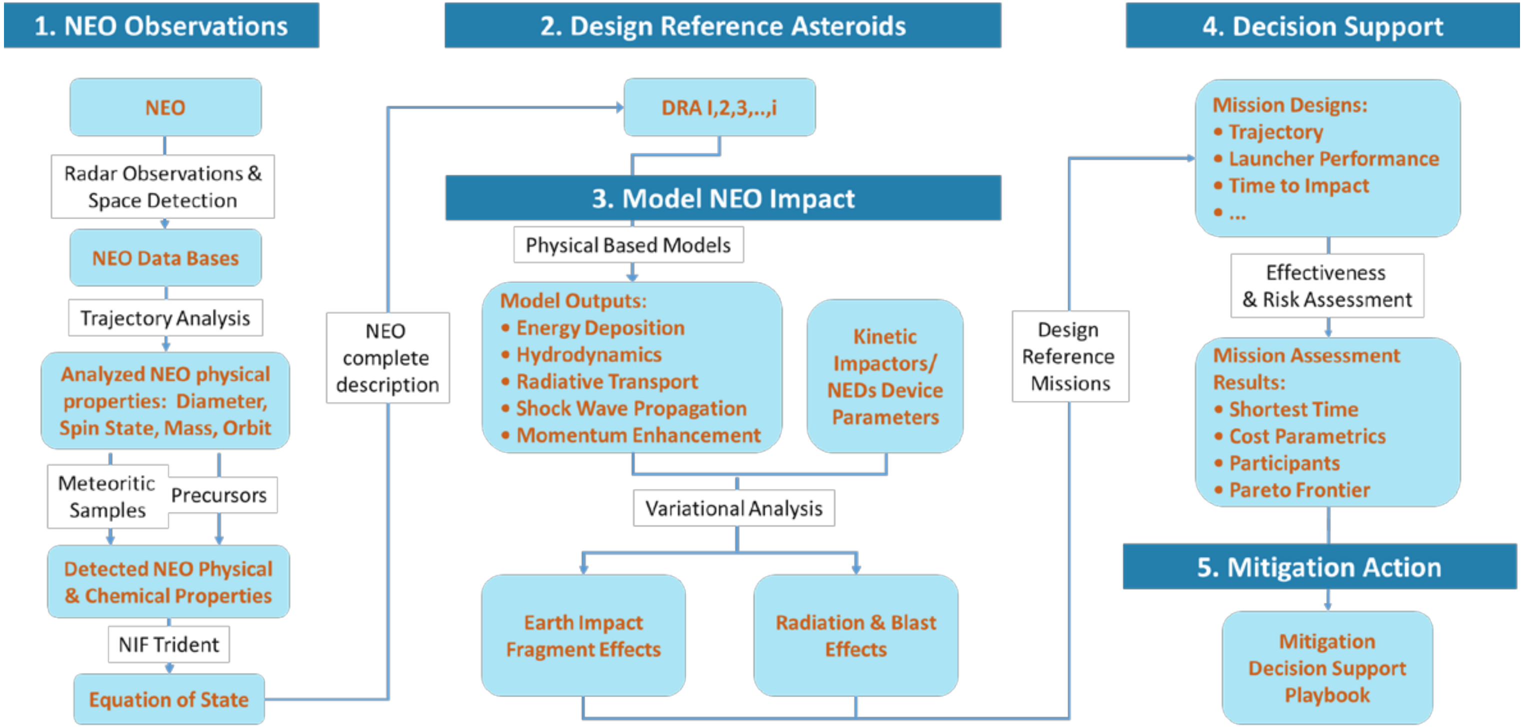


Architectural Framework

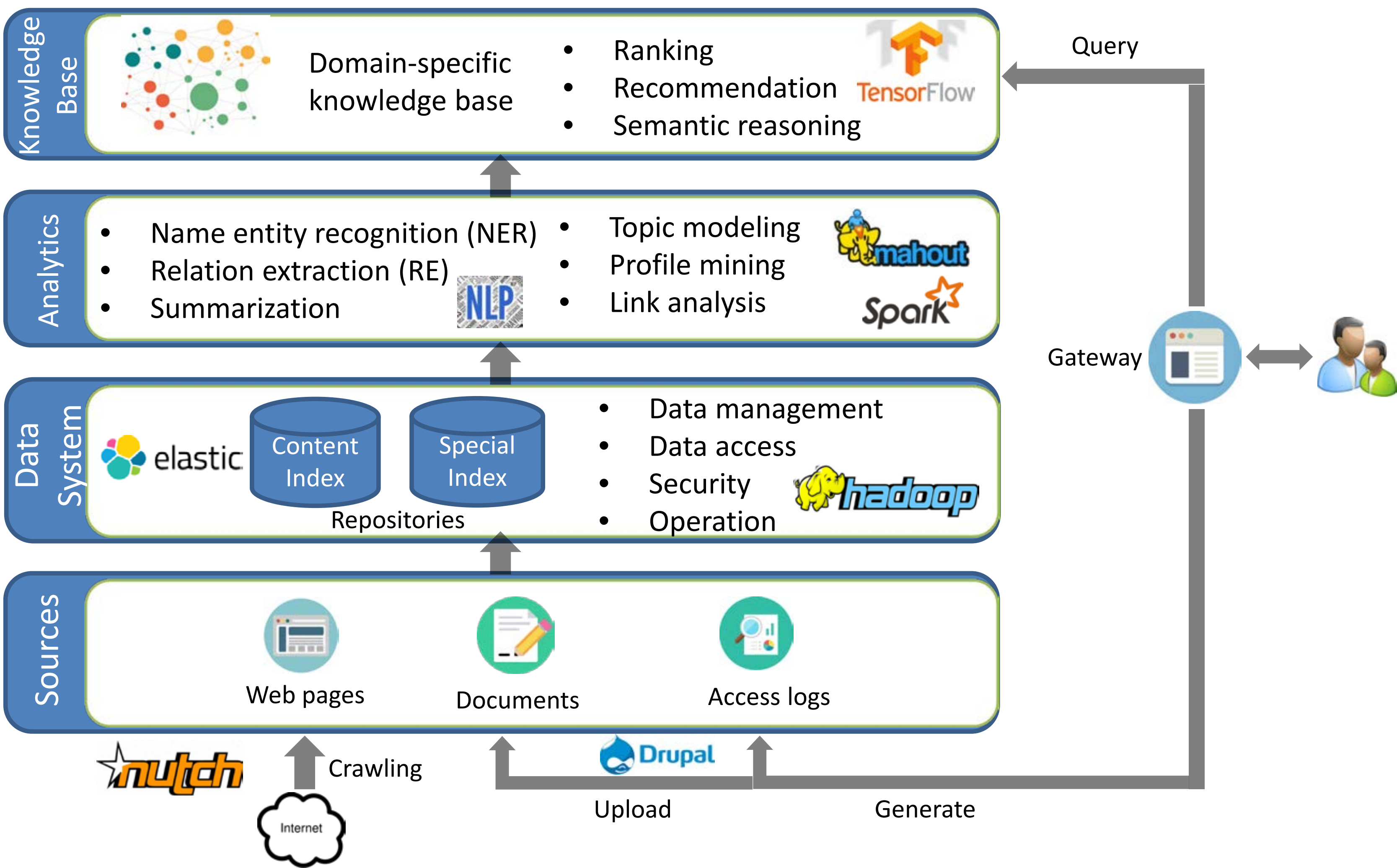
List is not exclusive



Information Flow

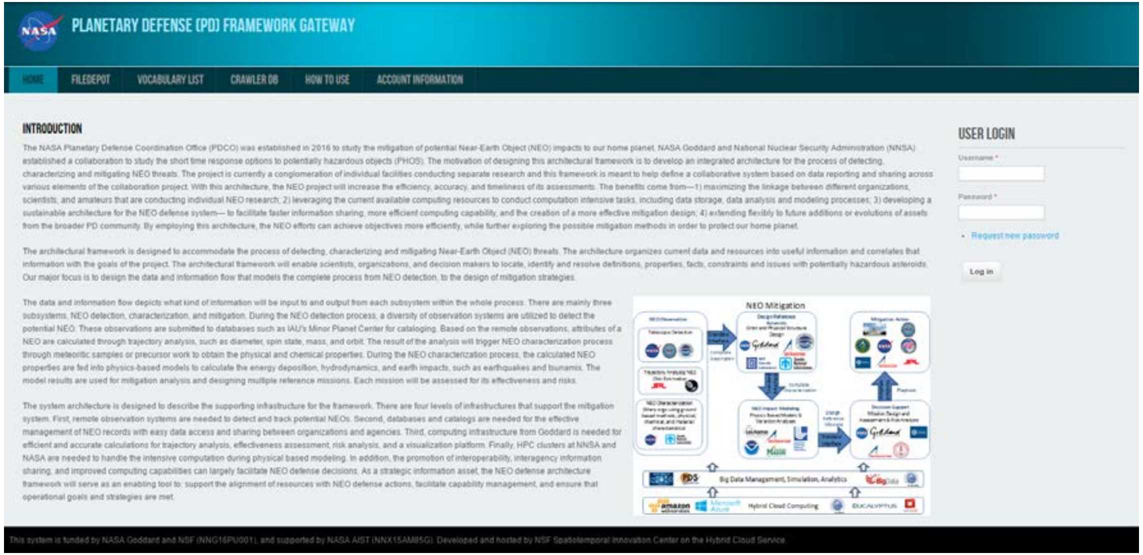


Knowledge Discovery & Usage Framework



Planetary defense (PD) Framework Gateway

- Web Portal: http://pd.cloud.gmu.edu/
- User management, document archiving, vocabulary editing web crawling, search engine

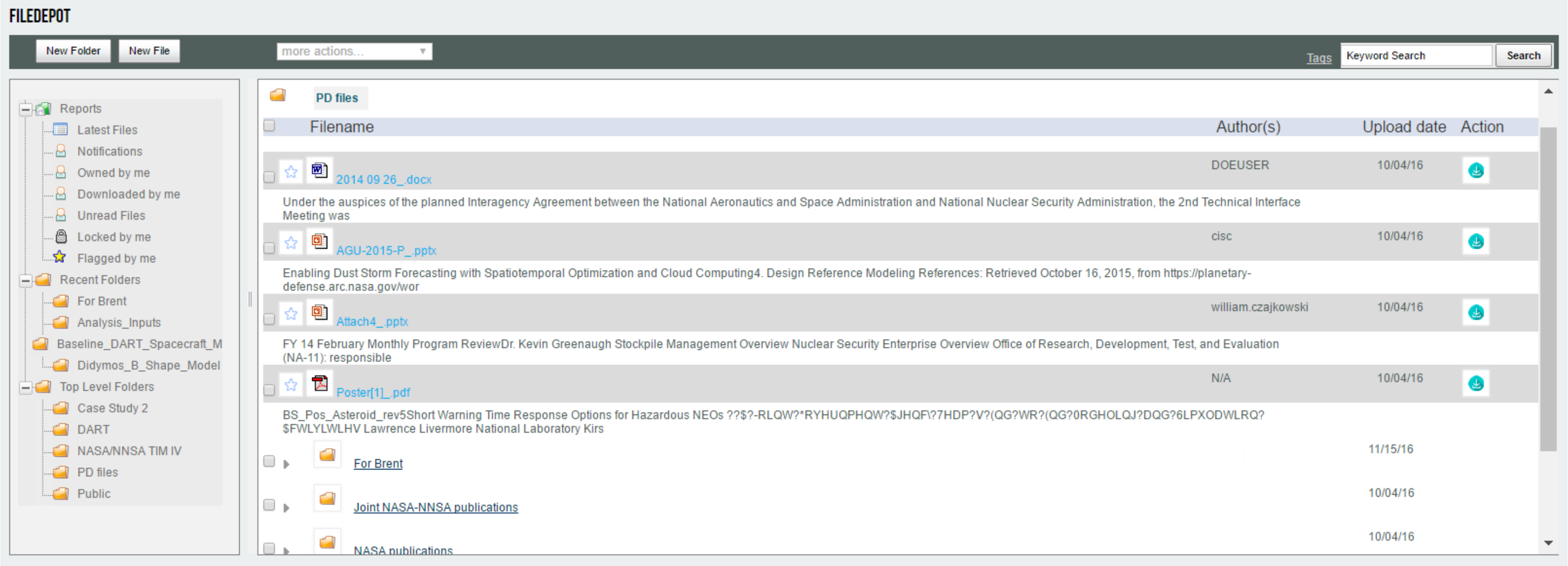


User management

- User roles: Administer, authenticated user, anonymous user
- Manage access control with permissions and user roles
- Assign permissions and roles to users
- Ban an IP address - The Ban module allows administrators to ban visits to their site from individual or a range of IP addresses.

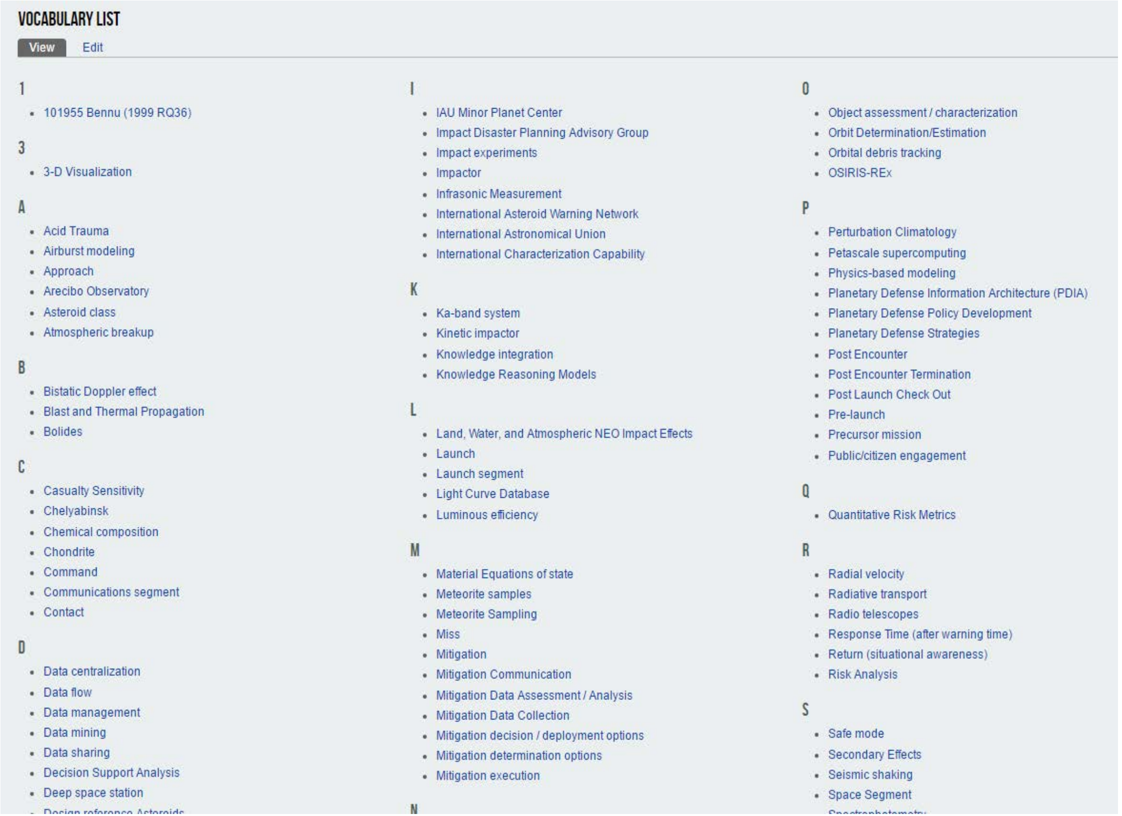
FileDepot Module: File/document Management

- Create folders or upload new files
- More actions



Vocabulary editing module

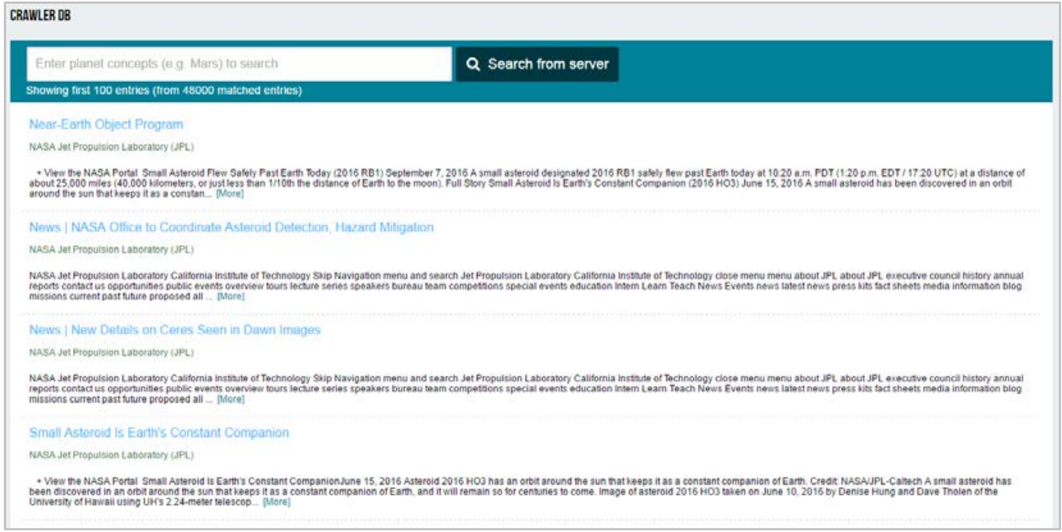
- 130+ concepts
- Create and edit landing page for each concept
- Different user roles have different permissions



Web crawling module

- Nutch: Open Source web crawler
- Store them in Elasticsearch (full-text search engine)
- 5 seed URLs
- Similarity between page vs. vocab list
- Baseline

```
controller.addSeed("http://neo.jpl.nasa.gov/");
controller.addSeed("http://global.jaxa.jp/");
controller.addSeed("http://neo.ssa.esa.int/");
controller.addSeed("http://neocam.ipac.caltech.edu/");
controller.addSeed("http://www.minorplanetcenter.net/iau/mpc.html");
```



Ongoing research

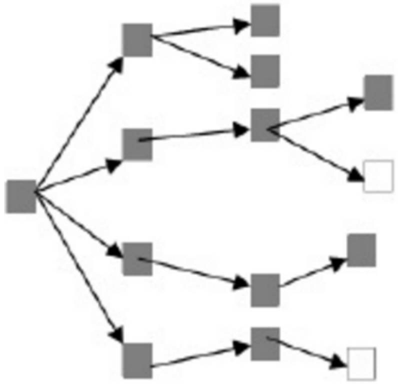
- Domain specific crawling
- Knowledge extraction from plain text

Domain specific crawling

Simplest approach: filter web pages using a keyword list (e.g. NEO, asteroid, Bennu, …) composed by domain experts.

Problems:

- Expensive
- Difficult to exhaust
- Difficult to assign weights to different keywords
- Treat all web pages equally (a page on NASA website and a random one)



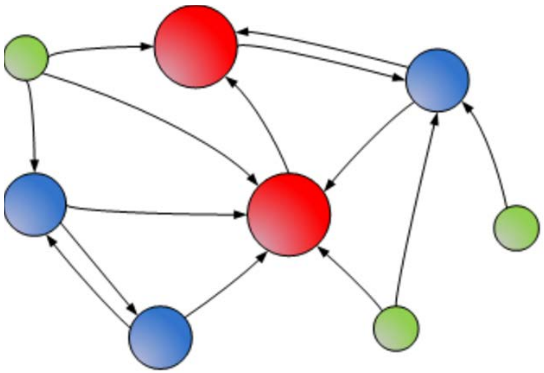
Distribution of relevant pages in blue

Image source: http://www.seminarsonly.com/computer%20science/focused-web-crawling-for-e-learning-content.php

Domain specific crawling

Existing tools in Open Source crawler (e.g. Nutch):

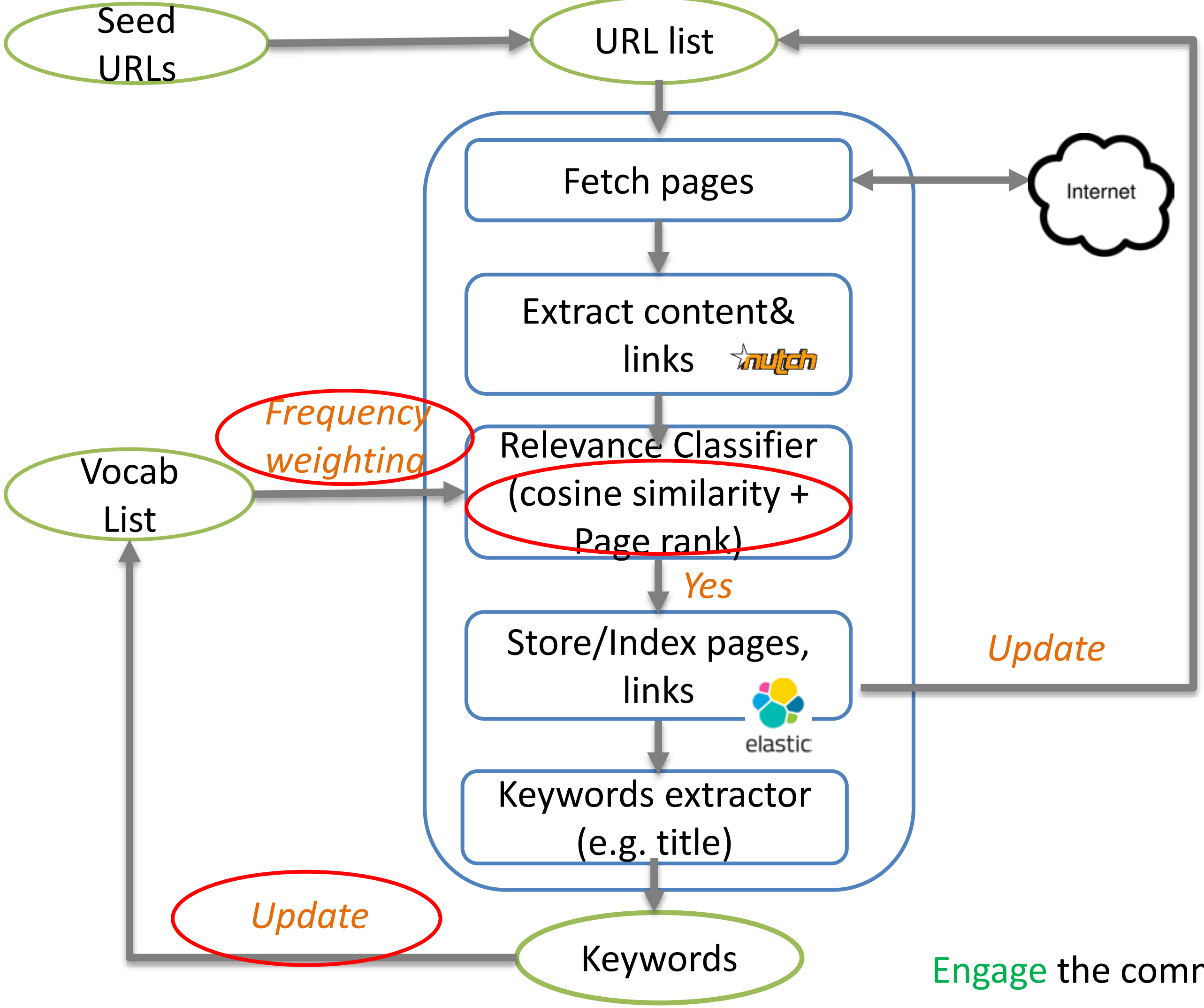
- Link-based
  - Scoring links (OPIC, PageRank scoring)
  - Breadth first or Depth first crawl



- Content-based
  - URL, mimetype filter
  - Cosine Similarity scoring filter (what we are using)
  - Naive Bayes parse filter

Image source: https://en.wikipedia.org/wiki/PageRank

Proposed method



- Combine content and link-based scoring to boost the authoritative and relevant web pages
- Dynamically update/grow the vocab using info (e.g. title) from the web pages
- Weight keywords based on frequency clustering (i.e. more frequently seen terms have more weights)

Engage the community to help with the evaluation

Knowledge extraction from plain text

- Goal: Extract structured information from unstructured web pages and user uploaded documents

- **Relation extraction** in NLP: finding semantic triples (*SPO*) from sentences

Predicate

The UV Index is a measure of the intensity of UV rays from the Sun.

Subject *Object*

- Pattern-based, supervised, semi-supervised, and open information extraction

Relation extraction

Hand-written patterns

- “Y such as X”
- “such Y as X”
- “X or other Y”
- “Y including X”

- + Tend to be high-precision
- + Tailored to specific domains
- - Human patterns are often low-recall
- - Hard to be exhaustive

Open Information Extraction

- Recently published by Univ. of Washington
- Extract relations from the sentences with no training data, no list of relations (unsupervised)
- Self-learning process, syntactic and lexical/semantic patterns

```
The U.S. president Barack Obama gave his speech on Tuesday to thousands of people.
```

⇩

```
(Barack Obama, is the president of, the U.S.)
(Barack Obama, gave, his speech)
(Barack Obama, gave his speech, on Tuesday)
(Barack Obama, gave his speech, to thousands of people)
```

Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. Leveraging Linguistic Structure For Open Domain Information Extraction. In Proceedings of the *Association of Computational Linguistics (ACL),* 2015.

Open Information Extraction

the GHRSST	is	a truly international project with over $18 Million USToday
Jason-1	has	a repeat period of approximately 10 days with 254 passes per cycle
Jason-3	is	capable of measuring significant wave height, sigma naught (sigma0), dry and wet tropos
The Aquarius instrument	has	3 radiometer beams in push-broom alignment with footprint resolutions of 76 km
Jason-3	has	a repeat period of approximately 10 days with 254 passes per cycle
Jason-1	is	capable of measuring significant wave height, sigma0, dry and wet troposphere and ionos
Level-2 data	refer	to monthly estimates of spherical harmonic coefficients of the Earth gravity field
no downlink signal	was detected	At the beginning of the next contact at 0249 UTC
sensors	included	a CTD at the near-surface and another at 6 m depthFor SPURS-1

- Some are reasonable, some are noise
- Working on reducing noise/identifying reasonable results

Conclusion and Next Steps

- The proposed architecture framework benefits the PD community by
  - Providing discovery and easy access to the knowledge and expert opinion within the project team
  - Maximizing the linkage between different organizations, scientists, engineers, decision makers, and citizens
- Next steps
  - Develop a knowledge base & search ranking for NEO mitigation resources
  - Investigate a knowledge reasoning model for potential mitigation by assimilating existing scenarios
  - Build a 4D visualization tool based on new datasets and existing tools

References

- Agichtein, E., Brill, E. and Dumais, S., 2006. Improving web search ranking by incorporating user behavior information. ed. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 19-26.
- AlJadda, K., *et al.*, 2014. Crowdsourced query augmentation through semantic discovery of domain-specific jargon. ed. *Big Data (Big Data), 2014 IEEE International Conference on*, 808-815.
- Auer, S., *et al.*, 2007. Dbpedia: A nucleus for a web of open data. *The semantic web.* Springer, 722-735.
- Bach, N. and Badaskar, S. 2007. A survey on relation extraction. *Language Technologies Institute, Carnegie Mellon University*.
- Banko, M., *et al.*, 2007. Open Information Extraction from the Web. ed. *IJCAI*, 2670-2676.
- Bargellini, P., *et al.*, 2013. Big Data from Space: Event Report. European Space Agency Publication.
- Battle, R. and Kolas, D. 2012. Enabling the geospatial semantic web with parliament and geosparql. *Semantic Web,* 3(4), 355-370.
- Cook, S., *et al.* 2011. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PloS one,* 6(8), e23610.
- Egenhofer, M. J., 2002. Toward the semantic geospatial web. ed. *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, 1-4.
- Graybeal, J., 2015. *Community Ontology Repository Prototype: Development* [online]. ESIP. Available from: http://testbed.esipfed.org/cor_prototype [Accessed 12/5 2016].
- Jiang, Y., *et al.* 2016a. A Comprehensive Approach to Discovering and Determining the Semantic Relationship among Geospatial Vocabularies - An Example with Oceanographic Data Discovery. *International Journal of Geographical Information Science*.
- Jiang, Y., *et al.* 2016b. Reconstructing Sessions from Data Discovery and Access Logs to Build a Semantic Knowledge Base for Improving Data Discovery. *ISPRS International Journal of Geo-Information,* 5(5), 54.
- Joachims, T., 2002. Optimizing search engines using clickthrough data. ed. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 133-142.
- Krisnadhi, A., *et al.*, 2015. The GeoLink modular oceanography ontology. ed. *International Semantic Web Conference*, 301-309.
- Lee, J.-G. and Kang, M. 2015. Geospatial big data: challenges and opportunities. *Big Data Research,* 2(2), 74-81.
- Li, W., Goodchild, M. F. and Raskin, R. 2014. Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. *International Journal of Digital Earth,* 7(1), 17-37.

- Manning, C. D., *et al.*, 2014. The Stanford CoreNLP Natural Language Processing Toolkit. ed. *ACL (System Demonstrations)*, 55-60.
- Miller, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM,* 38(11), 39-41.
- Nativi, S., *et al.* 2015. Big data challenges in building the global earth observation system of systems. *Environmental Modelling & Software,* 68, 1-26.
- Noy, N. F. and Musen, M. A., 2000. Algorithm and tool for automated ontology merging and alignment. ed. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-00). Available as SMI technical report SMI-2000-0831*.
- Pouchard, L., 2013. *ESIP Semantic Portal* [online]. ESIP. Available from: http://testbed.esipfed.org/node/1243 [Accessed 12/5 2016].
- Raskin, R. and Pan, M., 2003. Semantic web for earth and environmental terminology (sweet). ed. *Proc. of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*.
- Singhal, A. 2012. Introducing the knowledge graph: things, not strings. *Official google blog*.
- Soderland, S., *et al.* 2010. Adapting open information extraction to domain-specific relations. *AI magazine,* 31(3), 93-102.
- Srivastava, J., *et al.* 2000. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter,* 1(2), 12-23.
- Sun, A., Grishman, R. and Sekine, S., 2011. Semi-supervised relation extraction with large-scale word clustering. ed. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 521-529.
- Wu, L. and Brynjolfsson, E. 2013. The future of prediction: How Google searches foreshadow housing prices and sales. *Available at SSRN 2022293*.
- Yang C., H. Q., Li Z., Liu K., Hu F., 2016 2016a. Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*.
- Yang C., Y. M., Hu F., Jiang Y., Li Y. 2016b. Utilizing Cloud Computing to Address Big Geospatial Data Challenges. *Computers, Environment, and Urban Systems*.
- Zelenko, D., Aone, C. and Richardella, A. 2003. Kernel methods for relation extraction. *Journal of machine learning research,* 3(Feb), 1083-1106.
- Zhou, Y., *et al.*, 2008. Large-scale parallel collaborative filtering for the netflix prize. *Algorithmic Aspects in Information and Management.* Springer, 337-348.