

# **Semantic Web Research: Applications & Tools**

**Frank Olken**

Lawrence Berkeley National Laboratory &  
National Science Foundation

[folken@nsf.gov](mailto:folken@nsf.gov)

Presentation to:

CENDI/NFAIS/FLICC Workshop on  
**Semantic Web: Fact or Myth**

Nov. 17, 2009

Version 17

# Disclaimer

Opinions expressed in this talk are solely those of the author (Frank Olken) and do not necessarily reflect the positions, opinions, or policies of his employers (LBNL & NSF) or the U.S. Government.

# Semantic Web Research at NSF is Fragmented

- Semantic Web research funding at NSF is fragmented across many diverse programs in many different organizational units at NSF.
- Many of these programs do not view their research programs as “Semantic Web”.
- I think that it is unlikely that this situation will change soon.

# Semantic Web Research at NSF

- **Logic Programming** – CISE/CCF Software Eng.
- **Nonmonotonic Logic** – CISE/IIS/RI
- **Information Extraction** - CISE/IIS/RI
- **Ontology Tools, RDF Triple Stores, etc.** - CISE/IIS/III
- **Semantic Web Software Tools** – OCI/STCI
- **Domain specific ontologies** – various Directorates
- **Semantic Web Applications** – various Directorates, OCI/STCI, CISE/IIS/III
- **Data Integration, Resource Finding** – CISE/IIS/III & BIO/DBI/ABI

# This talk

- An overview of the semantic web research scene as seen by an NSF program manager
- Mostly discussion of U.S. civilian research activities:
  - Some NSF
  - Some NIH
  - Some European

Note: EU public funding of semantic web is much greater than U.S. public civilian funding of semantic web research

# What is wrong with keyword search?

- Search on Washington: which one?
  - Washington, DC, Washington state, George Washington, Denzel Washington, Washington = US Gov't, U. Washington, Washington Univ., ...
- Improve precision of search by better semantics

# Questions about Semantic Web

- What is it?
- What is it good for?
- What tools does one need?
- How does one get started?

# What is the semantic web?

- **Set of technologies:** semantic graphs, description logic, ...
- **Set of standards:** RDF, OWL, HTTP, RDFa, ...
- **Set of Tools:** Protege, Jena, Sesame, Semantic Media Wiki, ....
- **Set of artifacts:** DBpedia, Cyc, Linkeddata.org ...

# What is the semantic web good for?

- **Improved search** (now)
- **Improved classification** (now)
- **Facilitate controlled vocabulary development** (now)
- **Improved selective information dissemination** (now)
- **Information (data + schema) integration** (research)
- **Data mashups, visualization** (prototypes)
- **Automated web services synthesis** (research)
- **Expertise Finding** (some prototypes)
- **Automated Question Answering** (research)

# What became of the semantic web?

- **Originally:**
  - Manual semantic annotation of WWW pages
  - Converge on a single shared ontology
- **This has not (yet) happened.**
- **What is happening:**
  - Web of linked data
  - Information extraction from web pages, text
  - Many ontologies, ontology matching, ...
  - Steady growth of semantic web tools ...

# Semantic Web: logics or graphs?

- **Semantic web as a graph:**
  - RDF, RDFS, “triple stores”, SPARQL, ....
  - Graph theoretic models of taxonomies, partonomies, ontology mappings
- **Semantic web as logic:**
  - OWL, OWL DL, OWL 2, inference engines
  - Description Logic, Common logic ( First Order Logic)
  - Rules, Rule ML, Rule Interchange Format, rule engines
  - Logic programming (Prolog, etc.)
  - Non-monotonic logics, default logics, ...

# Logic(s) or Graph(s)

- Both approaches have their uses
- **Graph(s):**
  - Concrete, computationally tractable, easier to visualize, scalable, amenable to procedural programs, used for implementation, used for linked data on web
- **Logic(s)**
  - More abstract, computationally harder, more expressive, allows mechanized inference, formal semantics, declarative specification, used for concepts, rules, more complex ontologies, ....

# Graph Theoretic Semantics

- Taxonomies (is-a), Partonomies (part-of) are partial orders (directed acyclic graphs)
- Simple taxonomy = tree
- Multiple inheritance/facets = DAG
- Mappings between taxonomies should preserve relative order of nodes (concepts)
- See work by Cliff Joslyn @ PNNL.

# What is the semantic web?

- RDF, RDFS = semantic graphs
- RDF “Triple Stores”, SPARQL
- Linked Data = web of data
  - RDF + Use of URIs for universal addressing
- Logics: First Order Logic, Description Logic
  - OWL Full, OWL DL, OWL 2
- Rules = Rule ML, Rule Interchange Format, Rule Engines
- Inference engines

# What is the semantic web? (cont.)

- Standards for Semantic Markup of Web Content
  - RDF = semantic graphs
  - RDFS = RDF schema language
  - RDFa = RDF integrated in XHTML
  - Microformats = semantic hacks to XHTML, HTML
  - GRDDL = Rewrite rules for converting HTML microformats into RDF

# Semantic Web Tools

# What tools does one need?

- Ontology development tools
- Ontology matching tools
- Logic(s): Description Logic,
- RDF triple stores / SPARQL servers
- Rule markup languages, Rule interchange format
- Inference, rule engines
- Controlled natural languages
- RDF rendering/visualization tools

# RDF Triple Stores

- RDF = collection of triples, semantic graph, binary relational model
- Triple = (subject, predicate, object)
  - Subject, predicate are URIs
  - Objects are either URIs or literals
- Most triple stores are quad stores (to allow named graphs)
- Examples:
  - Jena, Sesame, Kowari/Mulgara, Oracle, Parliament (BBN), AllegroGraph (Franz), Abadi's work at Yale (NSF)
  - Some RDF triple stores have scaled > 1 Billion triples
  - Commonly support SPARQL query language

# Ontology Editors, IDEs

- Protege
  - From Stanford (Mark Musen, Natasha Noy, ...)
  - Supports OWL DL, ....
  - Pluggable RDF store, inference engines
  - Documentation, lots of users, plug-ins, ...
  - Open source, free, Java-based, ...
  - Training, user meetings (NIH)
- Others: Topbraid Composer, ...

# Faceted Search

- Organize taxonomy into multiple facets, not a single tree
- Use this to provide better browsing, search
- Apply to info retrieval, and product search
- Specify values for several facets: color, make, year, engine size, ...
- Examples:
  - **Flamenco**, by Marti Hearst (UC Berkeley)
  - <http://flamenco.berkeley.edu/> (NSF)
  - **RAVE** – statistical browser, Marchionini at UNC
  - <http://idl.ils.unc.edu/rave/>

# Inference Engines

- Mostly OWL DL (description logic)
- Plug-into RDF triple stores, mostly main memory
- Examples:
  - **Racer, Racer2** (Concordia, Hamburg, commercial)
  - **Pellet** (Clark & Parsia) (NSF)
  - **Jena2** (HP labs)
  - **F-Owl** (UMBC)
  - **Cerebra** (commercial), Oracle 11g (commercial)
  - **Silk** (from Vulcan, rules, logic programming, ...)

# Semantic Media Wiki Tool

- Semantically enabled wiki
- Simple syntax of a wiki
- Semantic annotations
- [http://semanticweb.org/wiki/Semantic\\_MediaWiki](http://semanticweb.org/wiki/Semantic_MediaWiki)
- SMW+ being developed by Vulcan, Ontoprise
  - <http://semanticweb.org/wiki/SMW%2B>
- Easy way to get started on semantic web authoring
- Also has been used (by NCI) for vocabulary development

# Silk

- **Semantic Inferencing on Large Knowledge**
- <http://silk.semwebcentral.org/>
- Supported by Vulcan, Inc.
- Contact: Benjamin Grosf
- Rules, defaults, logic programming, web scale
- Can support nearly all of Cyc (very large KB)
- Most ambitious semantic web rules project in the U.S.

# IKL

- Extension to First Order Logic
- Added named propositions – so you can talk about propositions
- Intended to support data integration, etc.
- Developed by Pat Hayes, Chris Menzel, John Sowa, et al.
- Reducible to Common Logic (NSF)
- Inference engines – not yet ...
- <http://www.ihmc.us:16080/users/phayes/silkie/Home.html>
- <http://www.ihmc.us/users/phayes/IKL/GUIDE/GUIDE.html>

# Spatial and Temporal Reasoning

- Various approaches:
  - Axiomatization in OWL + use OWL reasoner
    - Doable but slow ...
  - Special purpose predicates, indexing, etc.
    - **OWL** – Subrahmanian - temporal reasoning
    - **RDF** - Sheth - spatial reasoning (NSF)
    - **RDF** - AllegroGraph from Franz, Inc.

# RDF Output Formatting Tools

- RDF is not very readable
- Need the equivalent of XSLT, CSS to facilitate rendering of RDF into readable HTML, graphs, timelines, ...
- SIMILE Project
  - MIT
  - David Karger and students (NSF)
  - Exhibit, Fresnel, Timeline, Zotz, ...
  - <http://simile.mit.edu/>

# Information Extraction

- Use machine learning and NLP to extract “entities”, “facts”, taxonomy fragments from web pages.
- Exploit both language, web page structure (HTML & page layout), redundancy of many web pages, tables, lists, ....
- Bootstrap one's way to larger collections via semi-supervised learning
  - Andrew McCallum (U. Mass.) (NSF) Tom Mitchell (CMU), Yahoo, Google, Microsoft, Finin (UMBC) (NSF)
- **Major vehicle to populate the semantic web !!**

# Spreadsheets to RDF

- Spreadsheets are important source of legacy data.
- **RDF123** (Finin, Sachs @ UMBC) (NSF)
- **XLWrap** (Langeegger, Woss @Kepler Univ)
  - Supports SPARQL queries, multiple spreadsheets
- **Anzo for Excel** - Cambridge Semantics
- What is the difficulty?
  - Mapping multi-dimensional spreadsheets (> 2D)
    - Composite keys in column, row headings
    - Multiple spreadsheets (nested spreadsheets with composite keys in table headings)

# SPARQL to SQL translation

- Want to access legacy relational DBMS without converting entire DB to RDF Triple Store
- Requires knowledge of SQL schema, mappings from relation DB schema to RDF schema
- Multiple research efforts
- See **D2RQ**, **D2R** Map (Bizer @ Free Univ. of Berlin)
- Also recent work by Miranker (U.T. Austin)

# Zotero

- Firefox extension
- Collect, maintain, export bibliographic citations
- Imports many formats: End Note, Bibtex, ...
- Exports: RDF !!!
- Also, exports many bibliography formats
- Open source, from George Mason Univ.
- <http://www.zotero.org/>
- Zotz, <http://simile.mit.edu/wiki/Zotz>
  - Allows publishing of Zotero bibliographies via Exhibit

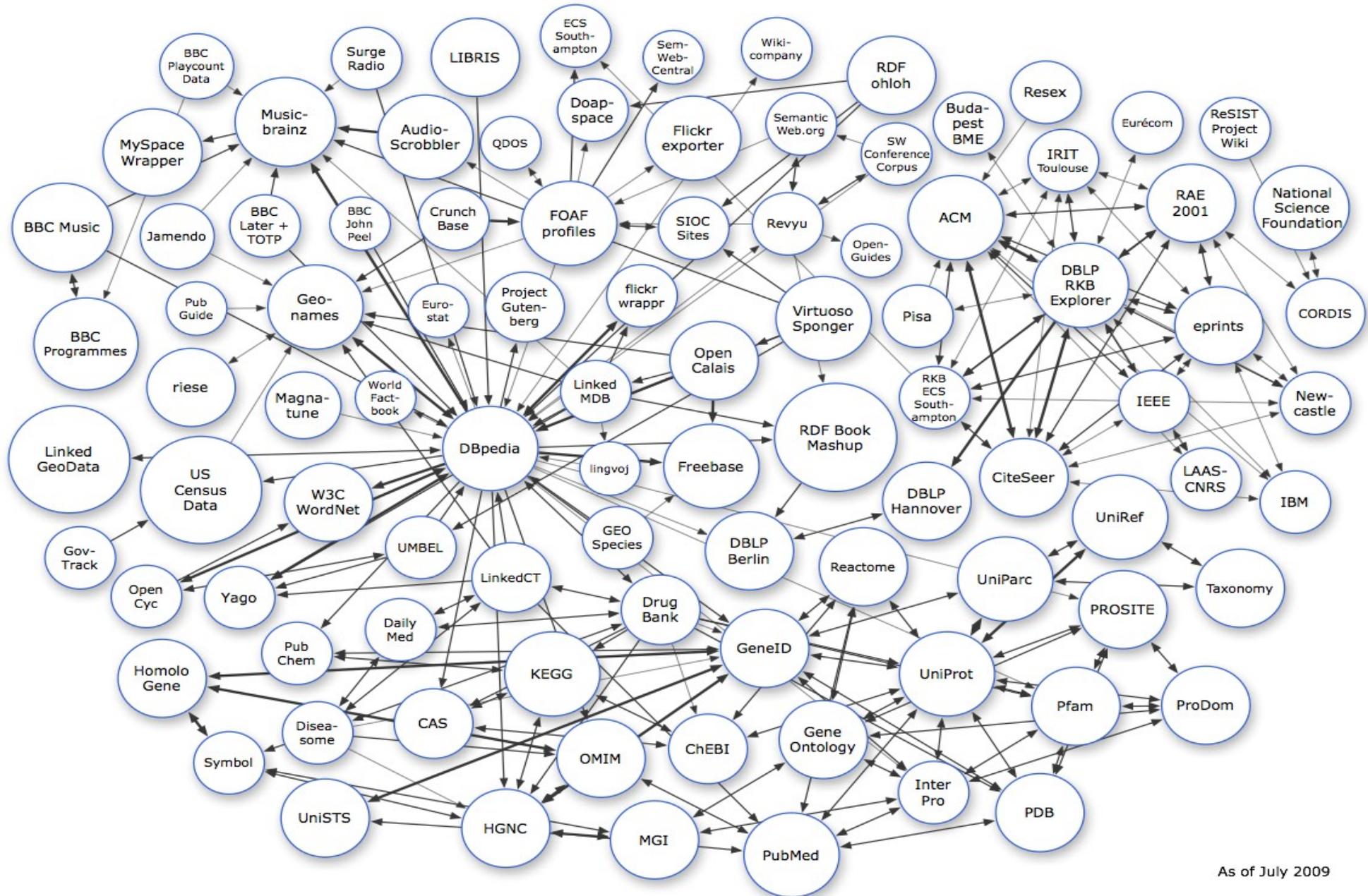
# Controlled Natural Language Tools

- Simplified form of English (or other languages)
- Simplified grammar (stilted but understandable)
- Easy to translate to/from formal logic
- Several efforts in UK, Europe
- Much easier to write or understand than formal logic
- Ease human interaction with software using formal logic
- See John Sowa: <http://www.jfsowa.com/talks/cnl4ss.pdf>

# Artifacts

- Cyc
  - Very large KB, 300K concepts, 2M propositions, ...
  - Developed by Doug Lenat, Cycorp.
  - Partially open source, partially in OWL, Silk?
- DBpedia
  - RDF encoding of Wikipedia
  - Very large KB, allows SPARQL queries
  - Good reference points for many concepts, persons, places, etc. for linked data

# Linked Data: Publishing RDF Data



- [linkeddata.org](http://linkeddata.org) - Fig. by Richard Cyganiak

# Getting Started on Semantic Web

- Use Protege (or equiv.) to convert your thesaurus, subject guide to OWL (not just SKOS); publish it on the web
- Publish your metadata for reports/datasets as RDF
- Experiment with Semantic Media Wiki
- Use RSS 1.0 (RDF based) syndication format for news
- Build a faceted browser to your report collection
- Build SPARQL/REST endpoints to your existing collections & DBMSs
- Use semi-supervised learning of your ontology

# Some Active Research Topics

- Probabilistic ontologies, reasoning, data, queries
  - Both in KB and DB communities
- Dealing with Inconsistent Data/Rules/Ontologies
- Parallel implementations of reasoners, triple stores
- Querying, reasoning across the web
- Modularization of ontologies, rule bases, ...
- Reasoning about numbers
- Spatial and temporal reasoning
- Ontology integration (matching)
- Non-monotonic reasoning (defaults)

# Conclusions

- Manual semantic annotation of web pages has not happened.
- Instead semantic web is being populated via:
  - Publishing linked data as RDF
  - Information extraction via semi-supervised ML+NLP
  - Creation of tools to access relational DBMSs and spreadsheets from SPARQL
- Semantic Web technology is useable now to improve retrieval, terminology development, data mashups, ...
- Do not need to convert all legacy DBMS to RDF triple stores

# Semantic Web Resources

# Semantic Web Web Sites

- W3C Semantic Web Activity
  - <http://www.w3.org/2001/sw/>
- Semweb Central
  - <http://semwebcentral.org/>
- Semantic Web Wiki
  - [http://semanticweb.org/wiki/Main\\_Page](http://semanticweb.org/wiki/Main_Page)
- Benjamin Grosf's Web Page
  - <http://www.mit.edu/~bgrosf/>
- Nodalities Magazine
  - <http://www.talis.com/nodalities/>

# Semantic Web Conferences

- Int'l Semantic Web Conference (**ISWC**)
- European Semantic Web Conference (**ESWC**)
- World Wide Web Conference (**WWW**)
- IEEE Int'l Conference on Semantic Computing (**ICSC**)
- Conference on Semantics for Health and Life Sciences (**CSHALS**)
- **OWLED** (OWL Experiences & Directions)
- **OIC** (Ontology for the Intelligence Community)

# Semantic Web Conferences (cont.)

- **Description Logic Conference**
- **Formal Ontologies for Information Systems (FOIS)**
- **Knowledge Representation Conference (KR)**
- **EKAW:** Intl Conf. On Knowledge Engineering and Knowledge Management, Knowledge Patterns (bienniel)
- **KCAP:** Conference on Knowledge Capture
- **Semantic Technology Conference** – trade show

# Semantic Web Journals

- **J. of Web Semantics** (Elsevier)
- **Applied Ontology** (IOS Press)
- **J. of Data Semantics** (Springer-Verlag LNCS)
- **Intl. J. on Semantic Web and Info. Systems** (IGI-Global)
- **Intl. J. of Semantic Computing** (World Scientific)
- **Nodalities** (STI)

# Semantic Web Standards Organizations

- **W3C (World Wide Web Consortium)**
  - RDF, RDF Schema, SPARQL
  - RIF (Rule Interchange Format)
  - OWL, OWL 2, OWL DL, OWL Full
- **ISO (International Standards Organization)**
  - Common Logic (CLIF)
  - ISO/IEC 11179 Metadata Registry Std
- **OMG (Object Management Group)**
  - Ontology Definition Metamodel (ODM)

# Books about the Semantic Web

- A Semantic Web Primer
- Semantic Web for the Working Ontologist
- Description Logic Handbook
- Foundations of Semantic Web Technologies
- Ontology Matching

# U.S. Semantic Web Research Groups

- Rennselaer Polytechnic Institute (Hendler, McGuinness, Fox)
- U. Maryland Baltimore County (Finin)
- Wright State Univ. (Sheth)
- MIT (Karger, Berners-Lee)
- U. Maryland College Park (Subrahmanian, Raschid)
- Yale (Abadi)
- SUNY Stonybrook (Kifer)
- Vulcan (Grosz, Greaves)

# U.S. Semantic Web Research Groups

- MITRE (Obrst)
- BBN (Dean)
- Bell Labs (Patel-Schneider)
- NIST (Neuhaus, Wallace)
- PNNL (Joslyn)

# Semantic Web Research Groups

- Canada
  - Toronto (Gruninger)
- Europe
  - DERI in Ireland
  - Karlsruhe, Germany
  - Vienna (STI)
  - Italy (Guarino)
  - Many others

# Ontolog Forum

- Web site:
  - <http://ontolog.cim3.net/cgi-bin/wiki.pl/>
- Weekly (approx.) Teleconferences
  - Lectures, Panels, Webinars, ...
  - Units of Measure Ontology Std Working Group
  - Ontology Summits (annual, April, at NIST)
- Organized by Peter Yim

# National Center for Biomedical Ontologies

- Stanford (Mark Musen)
- Univ. of Buffalo (Barry Smith)
- Web sites
- Bioportal (biomedical ontologies)
- Training Courses

# Acknowledgements

- This work has been supported by a grant (NSF-0941163) from the National Science Foundation, CISE Directorate, IIS Division, III Program to the Lawrence Berkeley National Laboratory via the NSF Internal Research & Development (IRD) portion of the IPA award.

# Acknowledgements

- I would like to thank several persons for advice, assistance, and instruction on various topics related to the semantic web and comments on drafts of this talk:
  - Joel Sachs (UMBC)
  - Benjamin Grosf (Vulcan)
  - Eric Neumann (W3C Life Sciences Working Group)
  - Evren Sirin (Clark & Parsia)
  - Chris Welty (IBM)
  - Mark Musen and Natasha Noy (Stanford)
  - David Karger (MIT)
  - Cliff Joslyn (PNNL)
  - Tina Gheen (NSF) & Maria Zemankova (NSF)

# Contact Information

- **Frank Olken**

- National Science Foundation
- 4201 Wilson Blvd, Suite 1125
- Arlington, VA 22230
- [folken@nsf.gov](mailto:folken@nsf.gov)
- 703-292-7350
- <http://twitter.com/frankolken>