Implementing 'Big Open Data' in government through Open Collaboration
- case examples and possibilities

Leveraging OOR

Peter Yim <peter.yim@cim3.com>
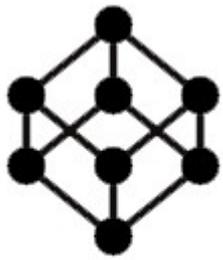
Mike Dean <mdean@bbn.com>

Leo Obrst <lobrst@mitre.org>

Ken Baclawski <kenb@ccs.neu.edu>

Presented at the

**Joint DATA.GOV - ONTOLOG "Big Open Data" Session**

17-May-2012

(v1.02)

1

Presentation Outline:

- *Introduction … and the ongoing dialog*
  - Peter Yim, Co-convener, ONTOLOG

- *The Ontology imperative for Big Data*
  - Leo Obrst, General Co-chair,
  Ontology Summit 2012: "Ontology for Big Systems"

- *Finding a Good Ontology* - Mike Dean,
  Co-convener, the OOR Initiative

- *A proposition on "Implementing 'Big Open Data' in government through Open Collaboration"*
 - Ken Baclawski, co-champion of the
  OOR Architecture and API effort
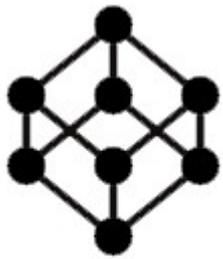
Among our earlier policy recommendations ...

- establish a subcommittee for "Ontology and Semantic Technology (OnST)" to foster inter-agency collaboration and align and coordinate their efforts.

- said OnST subcommittee to collaborate with the Digital Data IWG, Research Communities, Academia and like-minded International partners and communities to form a community of practice to mine collective intelligence and provide accelerated adoption

- support the effort (initiated since 2008) to build and maintain a global federated Open Ontology Repository (OOR)

- continue and accelerate the use of OnST to support the integration of medical data and Electronic Health Records as part of the mandate in our Healthcare Reform

- coordinate the creation of ontologies to strengthen and support the government open data initiative, and their incorporation into the above-mentioned open ontology repository

To: PCAST-NITRD (Jan-2011)
*From: Pat Hayes, Mark Musen & Peter Yim*
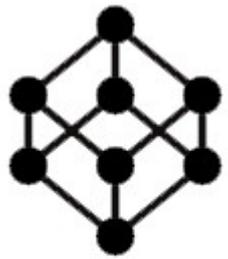
Ontology for Big Data

- We just finished with our Ontology Summit 2012 season - a 3-month rigorous discourse on "Ontology for Big Systems," where

- ~500 people – ontologists, system architects and engineers, and BigData/BigSystems stakeholders were engaged

- 28 virtual sessions, 99 presentations, over 1300 messages transpired

- Distilling that into a ten page Communique … and one conclusion we can draw, is:

  " … the unprecedented amounts of information and data are outstripping the capacities of current engineering practices and tools. Ontologies and ontological analysis are vital parts of a solution addressing the problems of architecting and engineering Big Systems and Big Data. …"
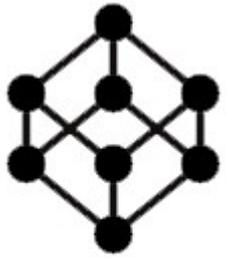
The Ontology Imperative

- Make explicit and accessible the implicit yet vital assumptions about the nature and structure of engineered systems and their components

- Help people better understand and disentangle the complexity of big engineered systems and their social, economic, and natural environment

- Enable integration among systems and data through semantic interoperability

- Allow humans to delegate more of the mundane processing and computing to machines (than was previously possible)

- Improve models and modeling, their adaptability and reuse, and resulting design

- Reduce development and operational costs

- Enhance decision support systems

- Aid in knowledge management and discovery

- Provide a basis for more adaptable systems
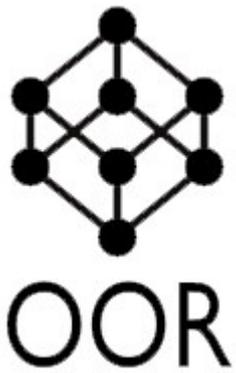
What is OOR?
the Open Ontology Repository Initiative

- **Community effort since January 2008**
- **Promote global use and sharing of ontologies**
  - *develop modular open source registry/repository software*
  - *Stand up one or more instances of OOR*
  - *promote best practices for ontology sharing and management*
- **Allow modular choices among**
  - Registry *functions (search, etc.)*
  - *Repository functions (persistence, versioning, etc.)*
  - *KR languages (OWL, Common Logic, etc.)*
  - *Gatekeeping policies (open, curated, market, etc.)*
  - *Intellectual Property Rights policies*
  - *Federation mechanisms (OOR, RSS, OMV, etc.)*
  - *Value-added services (alignment, translation, etc.)*
  - *…*
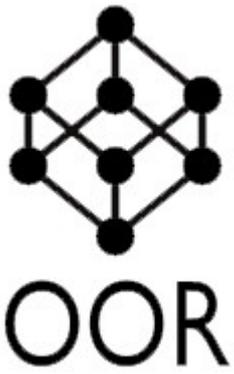
Finding a Good Ontology

At http://oor.net - access the OOR instances below,

BioPortal - http://bioportal.bioontology.org/ - the open repository for Biomedical Ontologies by the National Center for Biomedical Ontology (NCBO)

SOCoP-OOR - http://socop.oor.net - OOR instance by the Spatial Ontology Community of Practice (SOCoP)

MMI-ORR - http://mmisw.oor.net - the Ontology Registry & Repository for the Marine Metadata Interoperability (MMI) Project

ORNL-DAAC - http://ornl-daac.oor.net - the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC) for biogeochemical dynamics

COLORE - http://colore.oor.net - developing instance of the COmmon LOgic REpository (at University of Toronto)

Ontohub - http://ontohub.oor.net - the developing Ontohub instance for distributed ontologies (at University of Bremen)

OOR-dev - http://dev.oor.net - development instance for the OOR development team

OOR-test - http://test-00.oor.net - test instance for the OOR team

OOR-sandbox - http://sandbox.oor.net - OOR sandbox instance for shared open ontologies

Go to Homepage of the OOR Initiative - http://www.oor.net
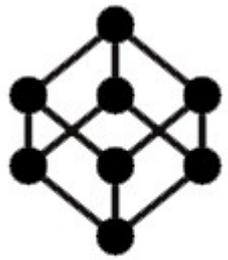
How does OOR plan to tackle it?

- OOR shall support evolutionary development
- OOR shall be scalable
- OOR shall support federation
- OOR shall provide services decoupled from core repository functionality, such as Semantic MediaWiki integration
- OOR shall support arbitrary ontology representation languages, including dataset schemas and data source ontologies
- OOR shall be ontologically driven
- OOR shall be platform independent
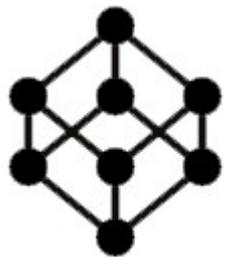
# Dealing with 400,000 Datasets? Here's one way to do it ...

- Don't stop at the data schema, vocabulary, ...

- Identify ontologies that are useful for the datasets, such as upper and super-domain ontologies
(like BFO, DOLCE, SUMO, PSL, UpperCYC, FEA-RMO, SWEET, ICOM, ...)

- Develop new domain ontologies as needed and host those ontologies on OOR

- Map dataset schemas and data source ontologies to domain, super-domain and upper ontologies

- Nurture the communities of practice that will allow the above work to be sustained

- Develop workflows and policies for ensuring the quality of ontologies and mappings
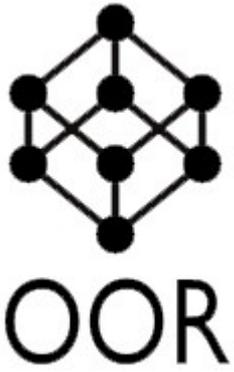
Next steps ...

- What do we (panelists in these two 'Big Open data' sessions) all have in common? … We are,
  - *Developing infrastructure with huge potential impact*
  - *Creating technology, harnessing content, and building trusted social networks of unprecedented value*
- This is hard work … but someone has to do it
- Accept that there are *no short cuts* to quality and to doing things right
- Team up, collaborate, and leverage each other's strengths and expertise
- Create the communities of practice and the ecosystem that allows these communities to thrive
- Let's keep this dialog going ...
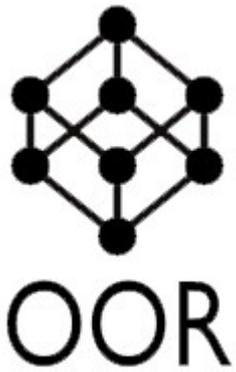
References

- The OntologySummit2012 "Ontology for Big Systems" Communique
  - http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2012_Communique

- Learnings from the past Ontology Summits
  - http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit

- The Open Ontology Repository Homepage
  - http://www.oor.net

- Accessing OOR instances – http://oor.net

- Join us …
  OOR - http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository#nid17YJ
  Ontolog - http://ontolog.cim3.net/cgi-bin/wiki.pl?WikiHomePage#nid1J