Semantic assets and challenges of ontologies management

Vasily.Bunakov <at> stfc.ac.uk
Science and Technology Facilities Council, United Kingdom

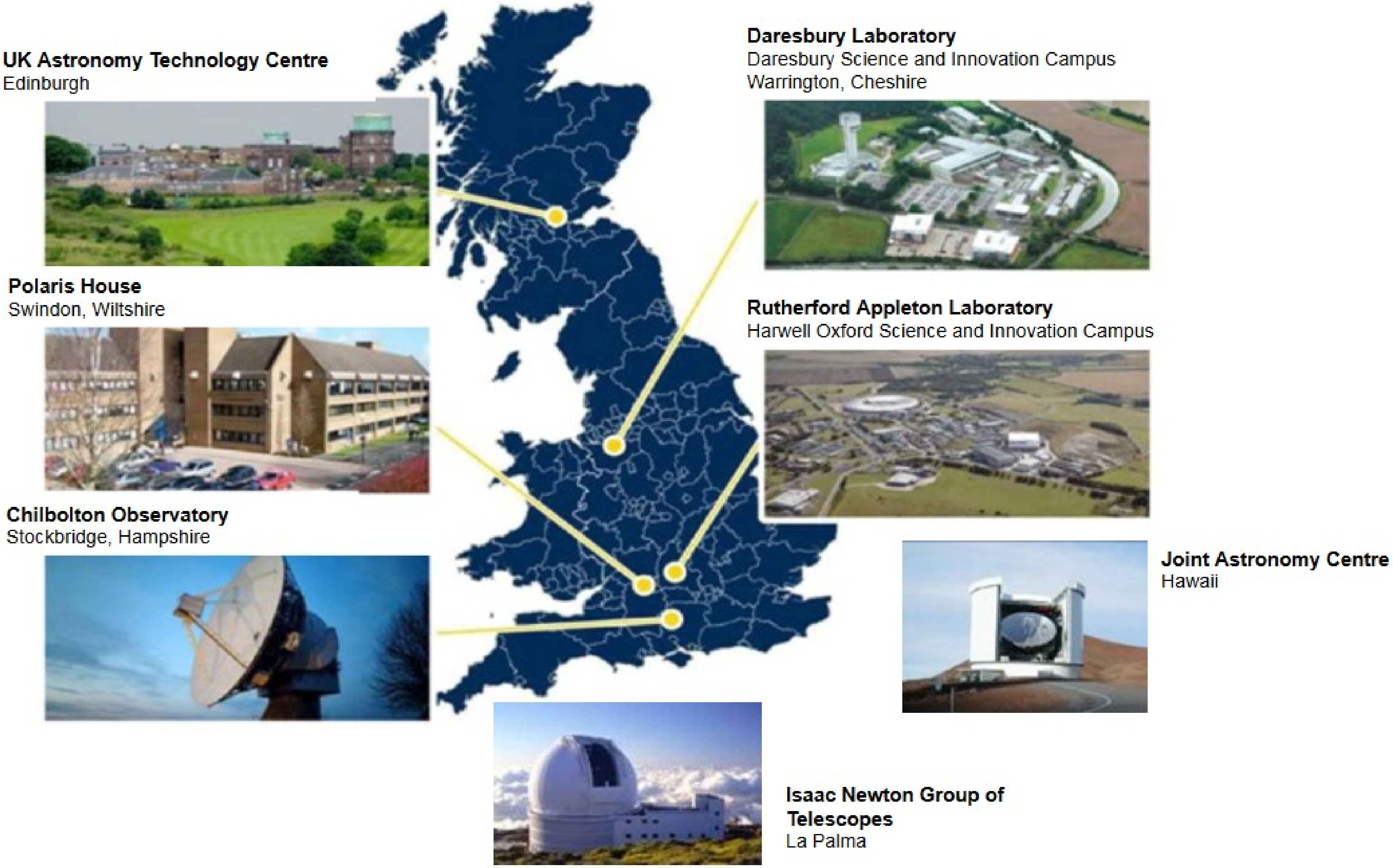
The EMMC IntOP2018 Workshop in Freiburg, 6-7 November 2018

TOC

- STFC and SCD background
- Semantic Assets for Materials Science Task Group
- Lessons from nano-foundries metadata design
- Lessons from elsewhere
- Suggestions on further communication

STFC and SCD background

STFC in a nutshell



UK Astronomy Technology Centre
Edinburgh

Polaris House
Swindon, Wiltshire

Chilbolton Observatory
Stockbridge, Hampshire

Daresbury Laboratory
Daresbury Science and Innovation Campus
Warrington, Cheshire

Rutherford Appleton Laboratory
Harwell Oxford Science and Innovation Campus

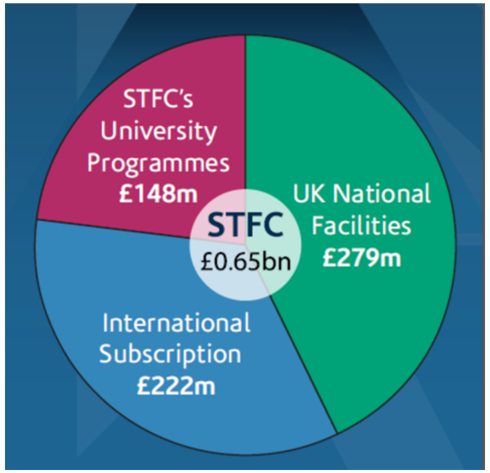
Joint Astronomy Centre
Hawaii

Isaac Newton Group of Telescopes
La Palma

Science & Technology Facilities Council

~ 1700 permanent staff
~ 7500 visitor scientists annually

2017



diamond

STFC Scientific Computing Department

Operate and develop IT infrastructure:

- High Performance Computing
- Petabyte data store
- CERN LHC Tier 1 hub
- Data management and data analysis solutions

This is where I come from

Do computational science:

- Biology and Life Sciences
- Engineering and Environment
- Computational Chemistry
- Theoretical and Computational Physics

See more at www.stfc.ac.uk/SCD

Physical Sciences Data Service

- Service to provide data resources to UK Chemistry and Materials Science Community
    - Extend a current service: http://cds.rsc.org/
    - Provide UK Academic access to commercial chemical databases
- University of Southampton and STFC taking over the service from Jan 2019
    - Initially transferring the current service from the Royal Society of Chemistry
- Plan to develop this as a Data Science platform
    - Develop it as a resource hub for Physical Sciences
    - Extend from Chemistry, to include Materials Science, Chemical Engineering and other related areas
    - More Open Science resources
    - Provide added value – common metadata, cross search, access to software, training
- Computed (simulated) datasets are identified as a possible territory for the service growth
- The advent of more machine-usable interfaces is foreseen
- Relation with NIST important

Recent EU projects with the STFC SCD contribution

- EUDAT – research data infrastructure
- EOSC – European Open Science Cloud
- VIMMP (well represented in this workshop)
- NFFA – Nanoscience Foundries and Fine Analysis
- FREYA – persistent identifiers in support of Open Science

We also contribute to a number of **RDA groups**, notably **Research data needs of the Photon and Neutron Science community IG** and **Vocabulary Services IG**

Semantic Assets for Materials Science Task Group

Semantic Assets for Materials Science Task Group

- Devised in the RDA Berlin plenary (April 2018), as a result of discussions between STFC and NIST
- Set up within the RDA Vocabularies Interoperability IG
- First online meeting in May 2018, followed by meetings in July and September
- Very open and inclusive group
- ~ 25 in the mailing list, ~ 10-12 a typical attendance
- Vasily Bunakov (STFC) and Zachary Trautt (NIST) co-chair

Semantic Assets Task Group scope

- Building an **inventory of existing semantic assets for Materials Science:** ontologies, vocabularies, controlled terms lists, metadata schemes . This can include not only vocabularies about materials per se but also cover adjacent topics, say instrumentation and chemistry, that are highly relevant for Materials community.
- Monitoring **technology for vocabularies building and vocabularies maintenance** / updates / curation in Materials domain
- Monitoring **use cases and actual practices for semantic assets application** in Materials domain. This includes using them in the actual IT services.
- Discussing **forms of representation / publishing for semantic assets**
- Discussing **interoperability between vocabularies**: a possibility for cross-walks or sensible links between terms from different vocabularies

Semantic Assets Task Group progress so far

- A good communication channel with representation from Europe and America; liaison with Japan / NIMS requires development
- First experiments with semantic assets registration using NIST platform http://schemas.nist.gov/
- Work on a common vocabulary started
- Potential for the F2F meeting in the RDA Plenary in Philadelphia (April 2019)
- Moving from the **RDA Vocabularies Interoperability IG** to the **RDA/CODATA Materials Data, Infrastructure & Interoperability IG** is possible

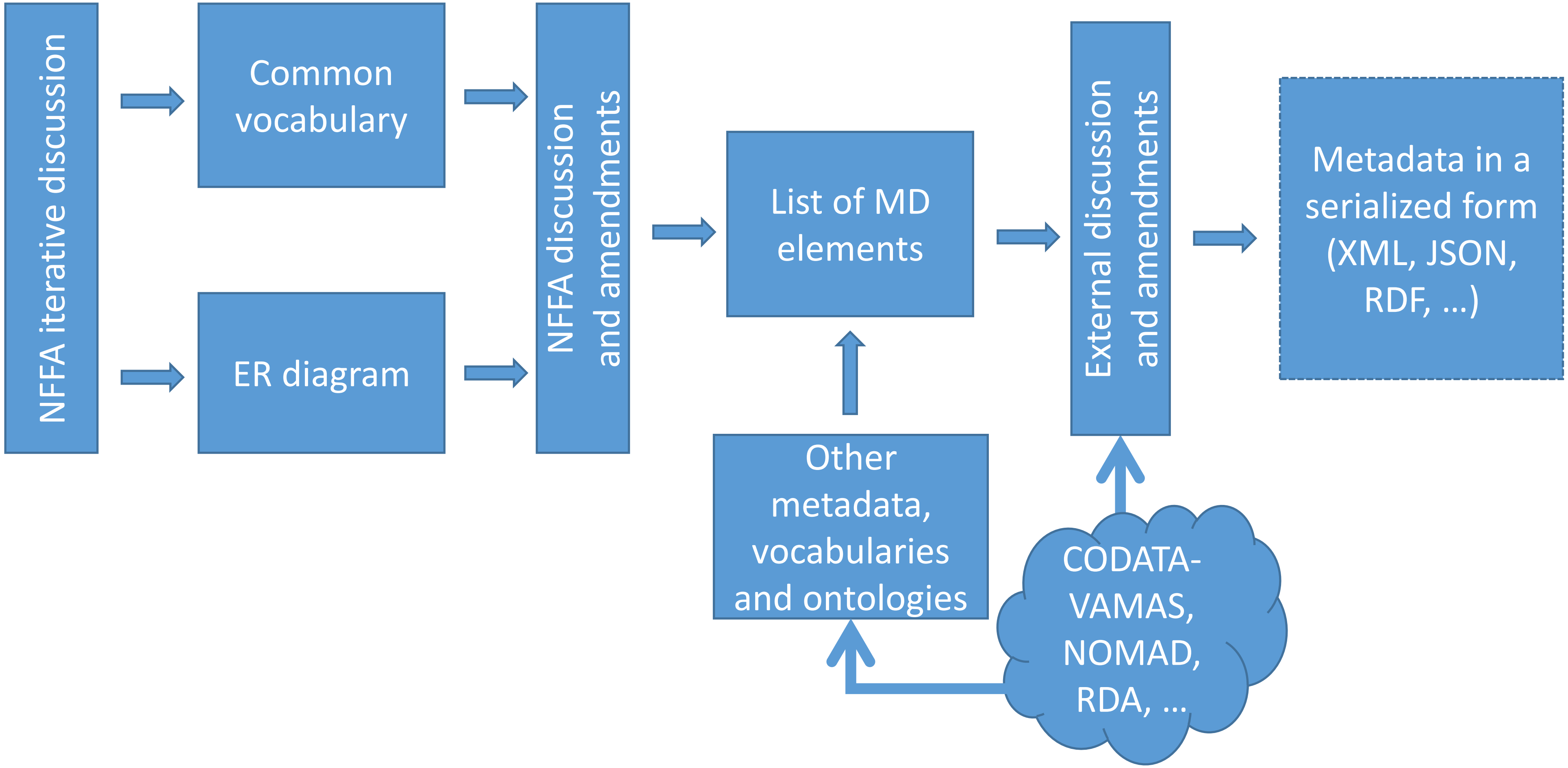
Lessons from NFFA metadata design

NFFA in a nutshell

- Is a Horizon 2020 project
- Gives access to distributed infrastructure for growth, nano-lithography, nano-characterization, theory and simulation and fine-analysis with synchrotron, FEL and neutron radiation sources
- “Virtual research enterprise” with proposals system and data management obligation

See more at www.nffa.eu

“What artefacts we produce” and “How we discuss them”: Stages of NFFA metadata design


NFFA iterative discussion
Common vocabulary
ER diagram
NFFA discussion and amendments
List of MD elements
Other metadata, vocabularies and ontologies
External discussion and amendments
CODATA-VAMAS, NOMAD, RDA, …
Metadata in a serialized form (XML, JSON, RDF, …)

An example of a semantic asset: A fragment of NFFA Common Vocabulary

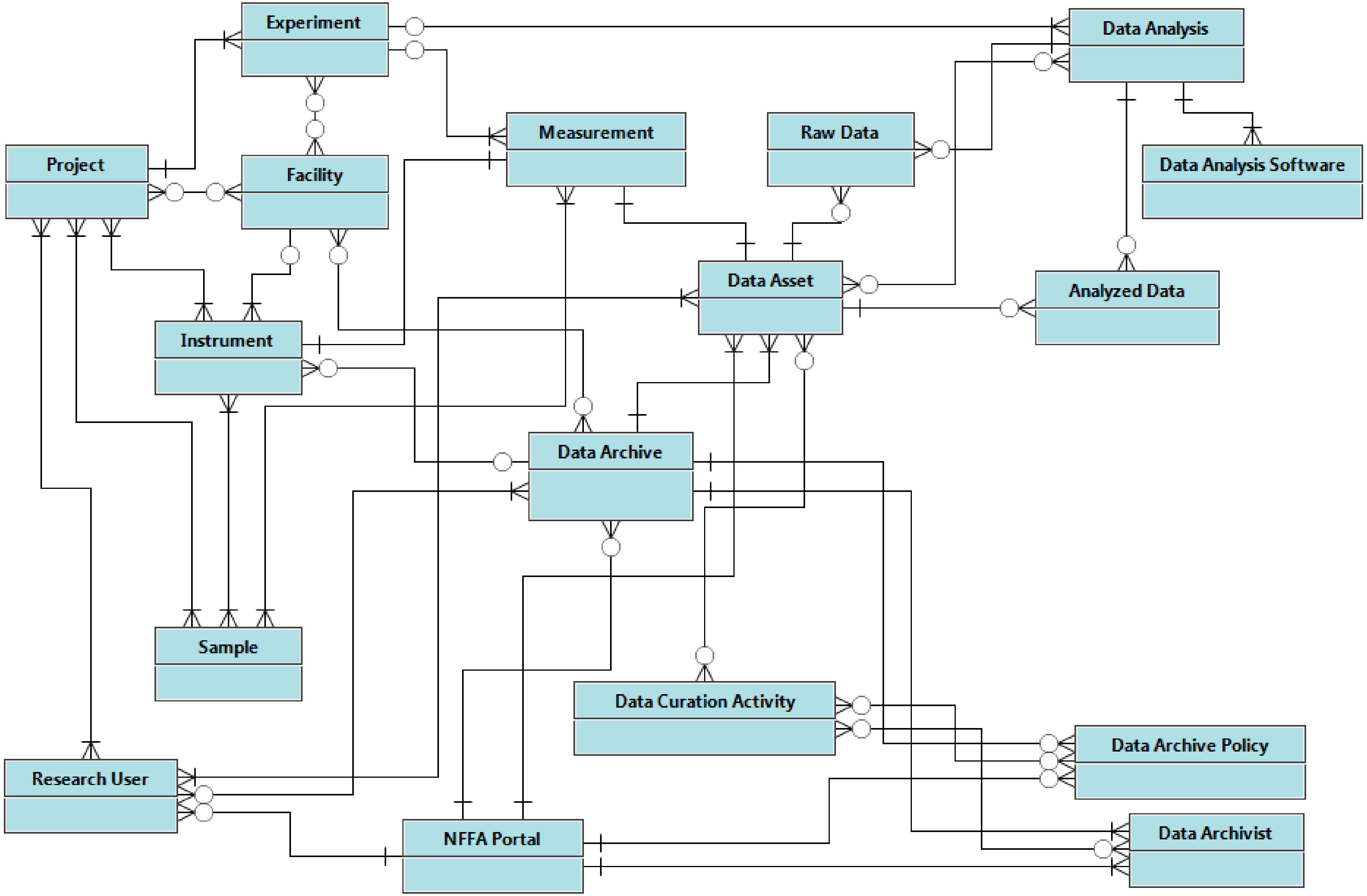
- **Research User.** A person, a group of them, or an institution (organization) who conduct Experiment on a nanoscience Facility using a nanoscience Instrument in order to collect and analyze Raw Data, or is interested in data collected or analyzed by other Research Users on the same or other Facilities.

- **Project.** An activity, or a series of activities performed by one or more Research Users on one or more Facilities using one or more Instruments for taking one or more Measurements of one or more Samples during one or more Experiments. Facility, Instrument, Measurement and Sample can refer to computer simulation environment.

- **Facility.** An institution (organization), or a division of it that operates one or more nanoscience Instruments for Research Users. For computer simulation, Facility can be a software platform that allows to order and manage computational experiments (so that the software platform serves the purpose of managing software modules that can be considered virtual Instruments).

- **Instrument.** Identifiable equipment (such as a device or a stand or a line) that allows conducting an independent nanoscience research, perhaps without involvement of other Instruments. Instrument is hosted by Facility and used by Research User. Instrument produces Raw Data in the course of Experiment. Instrument can be in fact a software for computer simulation (a software module or/and a particular configuration of it).

An example of a semantic asset:
ER diagram for NFFA metadata components



"No model is an island": Mapping and gap analysis exercise

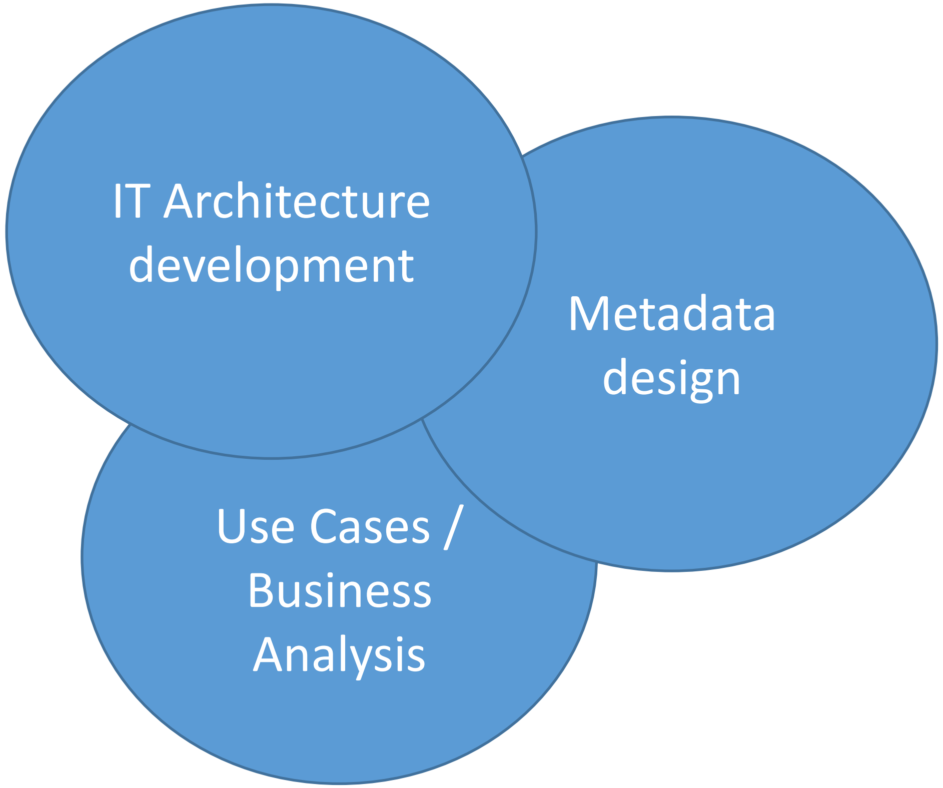
Concepts mapping

NFFA concept	CODATA-VAMAS concept	NOMAD concept
Experiment	Nano-object production steps	Series of software runs
Measurement	Nano-object testing steps	Software run
Sample	Nano-object or collection of objects	Input data
Data Asset		Output data

Models coverage / gaps

Nanotechnology aspect	NFFA model	CODATA-VAMAS model	NOMAD model
Nano-object (sample)	Conceptual	Detailed	Detailed
Computation	Detailed	Unaddressed	Detailed
Experiment lifecycle	Detailed	Conceptual	Conceptual
Data lifecycle	Detailed	Unaddressed	Conceptual

“Why do we do it at all”:
A place of metadata in a (virtual) Enterprise Architecture



Use Cases, IT Architecture and Metadata can be considered parts of a (virtual) Enterprise Architecture
See more about Enterprise Architecture at https://en.wikipedia.org/wiki/Enterprise_architecture

Lessons from semantic modelling beyond Materials Science

Ontology for finance

EDMCouncil

200+ organizations
7000+ professionals

www.edmcouncil.org

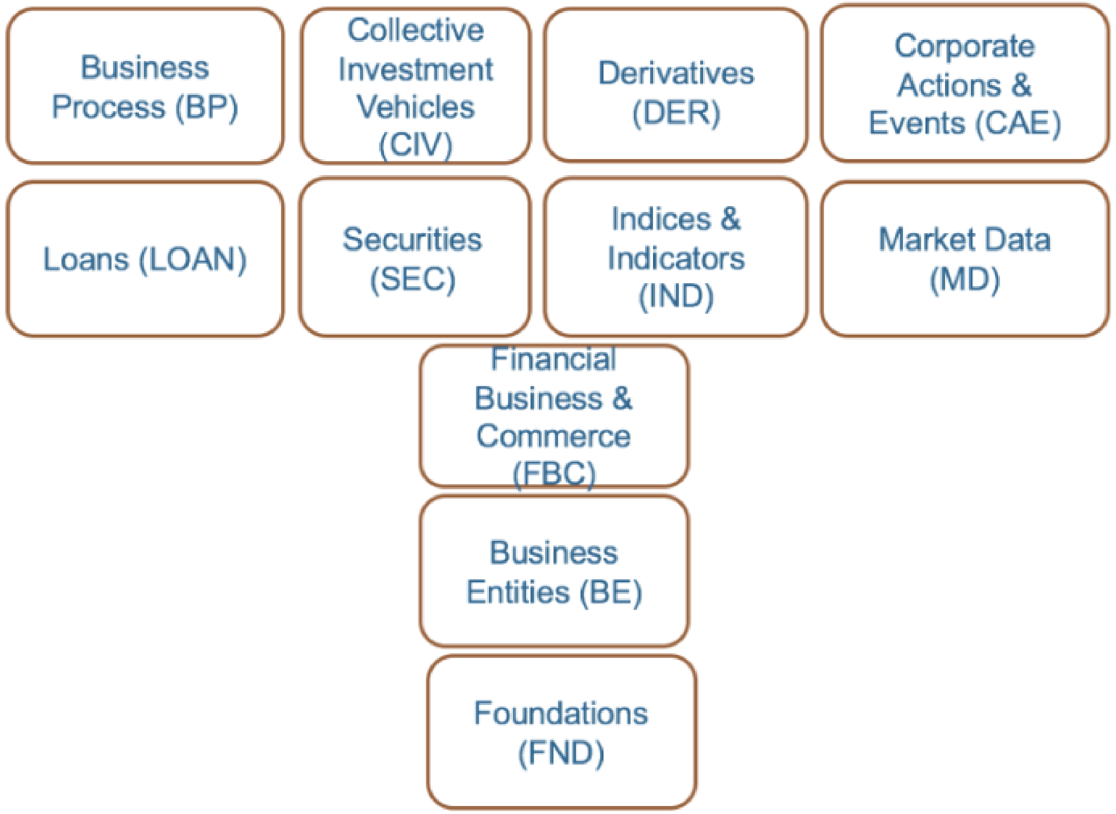
FIBO
Financial Industry Business Ontology

Business conceptual model of how all financial instruments, business entities and processes work in the financial industry

https://spec.edmcouncil.org/fibo/

FIBO is a well-governed project started circa 2010 and supported by a well-fed world-wide organization

Ontology for finance (continued): FIBO structure vs FIBO teams



Business Process (BP)

Collective Investment Vehicles (CIV)

Derivatives (DER)

Corporate Actions & Events (CAE)

Loans (LOAN)

Securities (SEC)

Indices & Indicators (IND)

Market Data (MD)

Financial Business & Commerce (FBC)

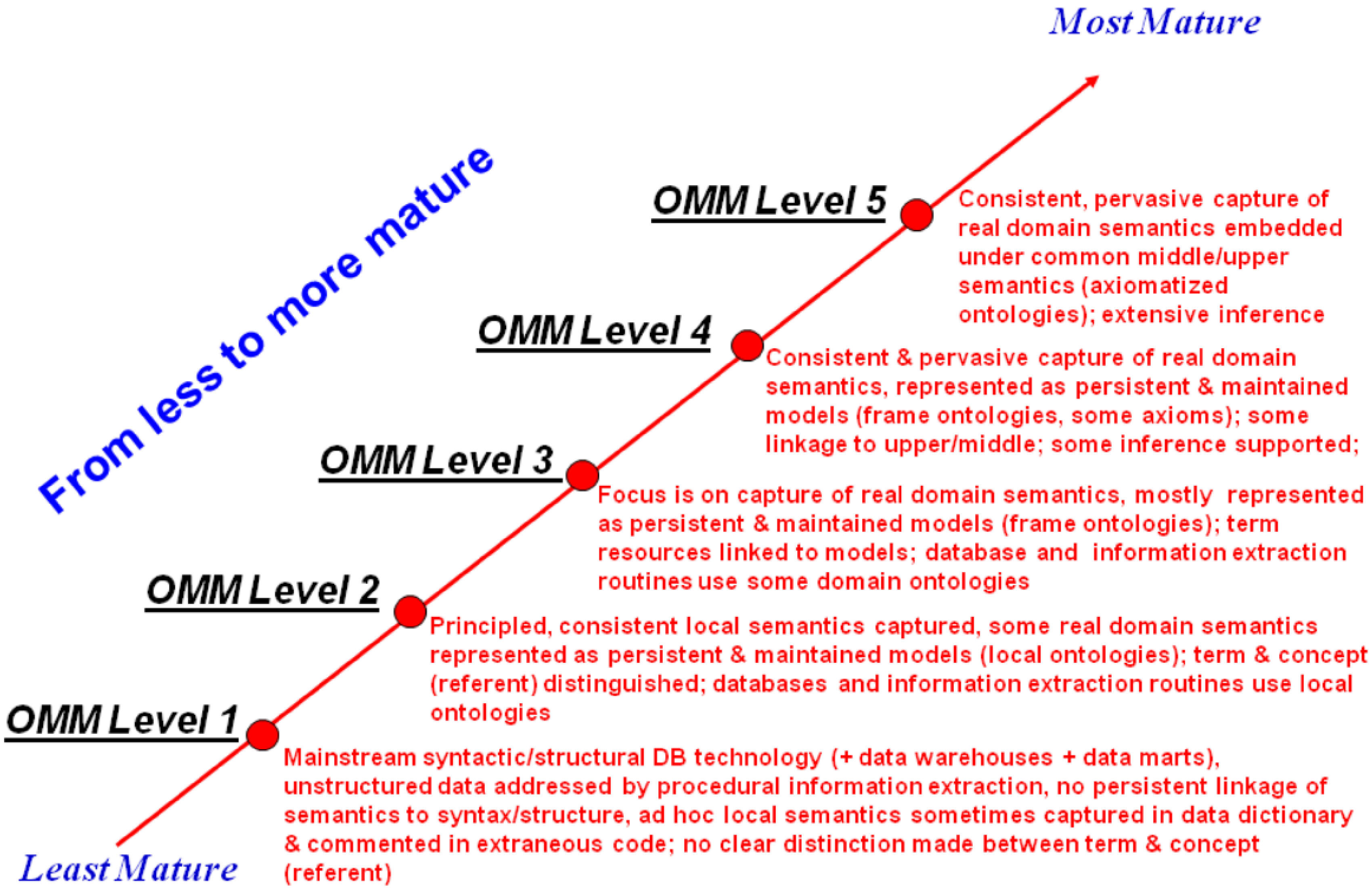
Business Entities (BE)

Foundations (FND)

- FIBO Leadership Team (FLT)
- FIBO Process Team (FPT)
- FIBO Proof-of-Concept Teams
- FIBO Foundations (FND)
- FIBO Business Entities (BE)
- FIBO Financial Business & Commerce (FBC)
- FIBO Indices and Indicators (IND)
- FIBO Securities & Equities (SEC)
- FIBO Derivatives (DER)

12 vendors are reported so far as having implemented FIBO in their IT solutions.
Not all parts of the model are currently covered by FIBO teams.

Ontology Maturity Model that informs FIBO development process



Most Mature

From less to more mature

OMM Level 5

Consistent, pervasive capture of real domain semantics embedded under common middle/upper semantics (axiomatized ontologies); extensive inference

OMM Level 4

Consistent & pervasive capture of real domain semantics, represented as persistent & maintained models (frame ontologies, some axioms); some linkage to upper/middle; some inference supported;

OMM Level 3

Focus is on capture of real domain semantics, mostly represented as persistent & maintained models (frame ontologies); term resources linked to models; database and information extraction routines use some domain ontologies

OMM Level 2

Principled, consistent local semantics captured, some real domain semantics represented as persistent & maintained models (local ontologies); term & concept (referent) distinguished; databases and information extraction routines use local ontologies

OMM Level 1

Mainstream syntactic/structural DB technology (+ data warehouses + data marts), unstructured data addressed by procedural information extraction, no persistent linkage of semantics to syntax/structure, ad hoc local semantics sometimes captured in data dictionary & commented in extraneous code; no clear distinction made between term & concept (referent)

Least Mature

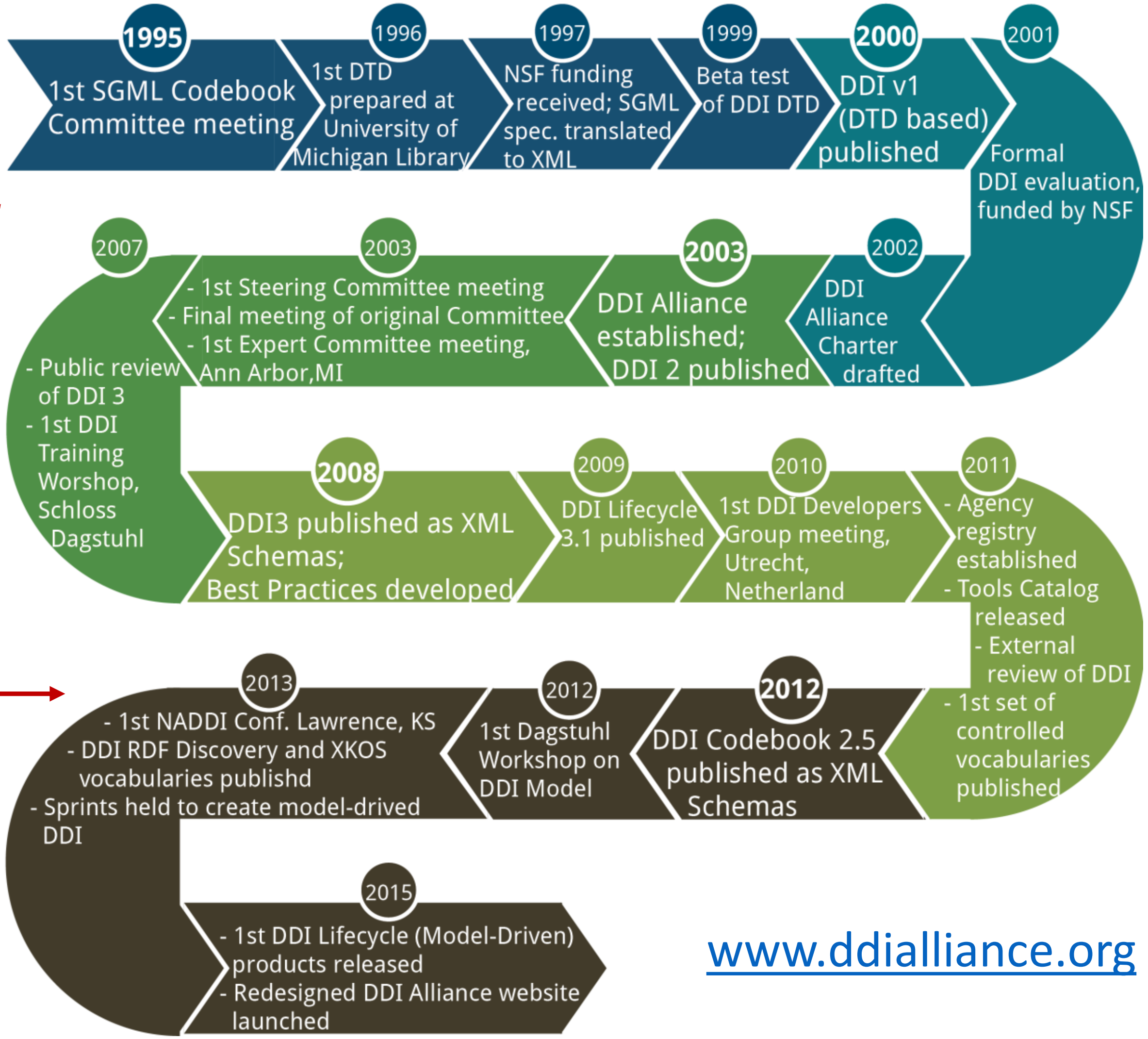
“The Ontology Maturity Model” by Leo Obrst, 2009 (inspired by CMM/ CMMI model for business processes maturity)

(a kind of) Ontology favoured by social science data archives

DATA DOCUMENTATION INITIATIVE

An international standard for describing surveys, questionnaires, statistical data files, and social sciences study-level information

It took 18 years from the first codification of terms to the first (incomplete) semantic representation. The official serialization is still XML Schema.



1995 1st SGML Codebook Committee meeting

1996 1st DTD prepared at University of Michigan Library

1997 NSF funding received; SGML spec. translated to XML

1999 Beta test of DDI DTD

2000 DDI v1 (DTD based) published

2001 Formal DDI evaluation, funded by NSF

2002 DDI Alliance Charter drafted

2003 DDI Alliance established; DDI 2 published

2003
- 1st Steering Committee meeting
- Final meeting of original Committee
- 1st Expert Committee meeting, Ann Arbor,MI

2007
- Public review of DDI 3
- 1st DDI Training Worshop, Schloss Dagstuhl

2008 DDI3 published as XML Schemas; Best Practices developed

2009 DDI Lifecycle 3.1 published

2010 1st DDI Developers Group meeting, Utrecht, Netherland

2011
- Agency registry established
- Tools Catalog released
- External review of DDI
- 1st set of controlled vocabularies published

2012 DDI Codebook 2.5 published as XML Schemas

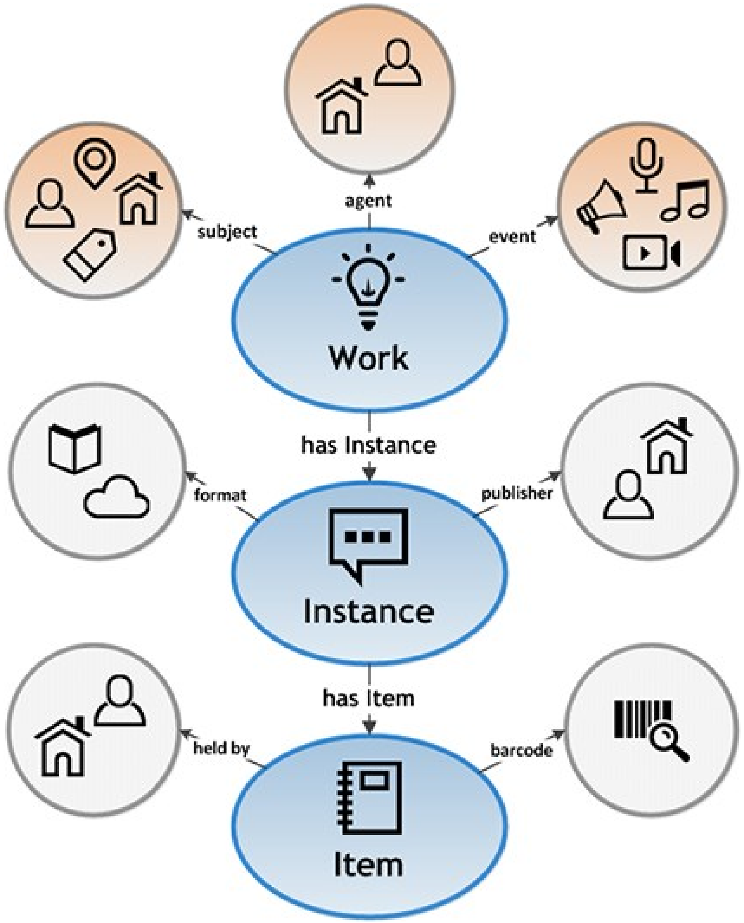
2012 1st Dagstuhl Workshop on DDI Model

2013
- 1st NADDI Conf. Lawrence, KS
- DDI RDF Discovery and XKOS vocabularies publishd
- Sprints held to create model-drived DDI

2015
- 1st DDI Lifecycle (Model-Driven) products released
- Redesigned DDI Alliance website launched

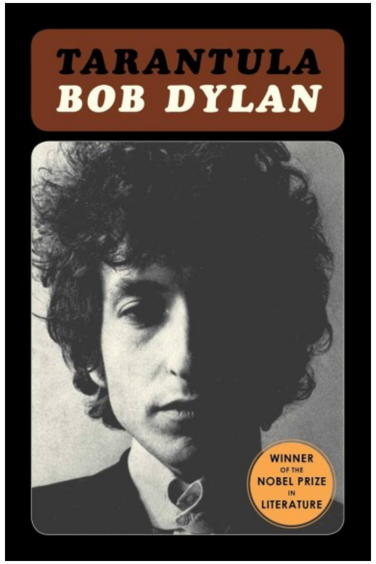
www.ddialliance.org

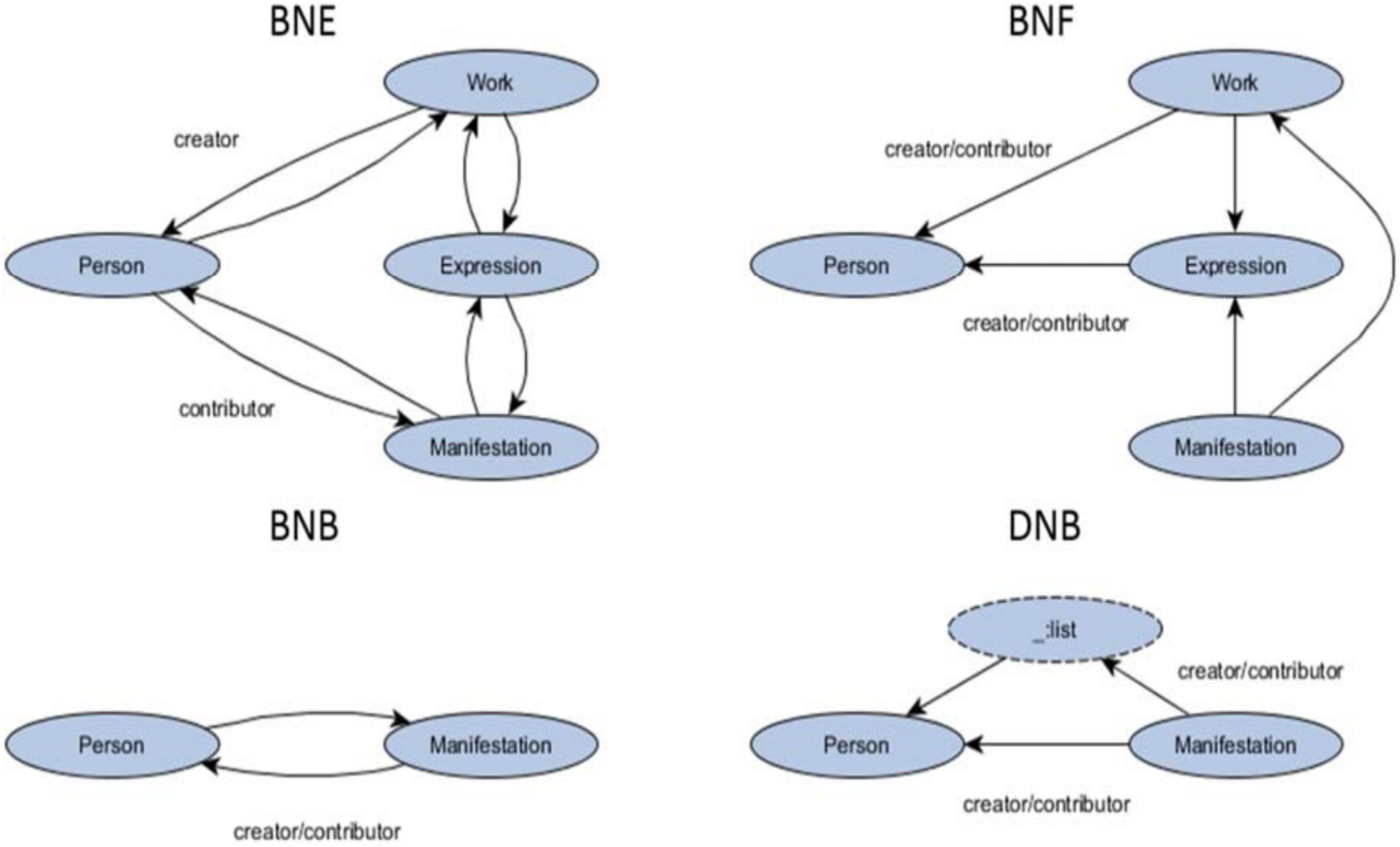
Ontology for bibliography (one of a few out there)



- 1960s: MARC Standards developed
- 1971: MARC become a national standard in the US
- 1973: MARC becomes an international standard
- 2002: library technologist Roy Tennant argued that "MARC Must Die", as it is used only within the library community, and designed to be a display, rather than a storage or retrieval format
- 2008: report from the Library of Congress wrote that MARC is "based on forty-year old techniques for data management and is out of step with programming styles of today"
- 2012: the Library of Congress announced that it had contracted with Zepheira, a data management company, to develop a linked data alternative to MARC
- 2012: the library released a draft of the new model, named BIBFRAME
- 2016: The Library of Congress released version 2.0 of BIBFRAME

The actual experiment of transforming MARC records to Linked Data by four national libraries)*





)* As presented in MTSR 2018 conference by Prof. Christos Papatheodorou, Ionian University, Corfu, Greece
Detailed description of experiment: Tallerås, K. (2017). Quality of linked bibliographic data: The models, vocabularies, and links of data sets published by four national libraries. Journal of Library Metadata, 17(2), 126–155. https://doi.org/10.1080/19386389.2017.1355166

Linked Data by 4 national libraries continued (something about semantics and interoperability)

- 3 of 1,141 unique property and class terms are used by all 4 libraries (owl:sameAs, rdf:type, and dct:language)
- 13 terms by (sets of) 3 libraries
- 34 terms by (sets of) 2 libraries

Why these three?

Set	***Triples***	***Entities***	***Data-level constants***
BNB	104,139,477	10,126,344	52,671,707
BNE	71,199,698	5,763,188	56,681,387
BNF	304,587,809	30,671,400	192,224,487
DNB	329,261,459	32,673,901	250,613,437
Average	202,297,111	19,808,708	138,047,754

Picture credits: “Three stones of wisdom” by http://livertising.net/blog/2013/three-stones-of-wisdom-livertising-exam-concepts/

Ontologies for biology)*

Rationale for ontologies repository

- Ontologies can be complex
- Ontologies can be big
- Ontologies can change

Ontology repository use cases

- Search for terms
- Querying the hierarchy
- Querying across relations

https://www.ebi.ac.uk/ols/index

(as per 1 November 2018)

216 ontologies
5,526,032 terms
19,119 properties

)* Simon Jupp (EU Bioinformatics Institute, Cambridge, UK). Building a repository of biomedical ontologies with Neo4j. https://www.slideshare.net/thesimonjupp/building-a-repository-of-biomedical-ontologies-with-neo4j

Semantic modelling and technology with no RDF involved

Flexible MDM (Master Data Management) with graph database: https://neo4j.com/case-studies/schleich/

Picture credits: https://www.ebay.co.uk/usr/bargain-vapes

We may have learned something about semantic interoperability…

- Ontologies / semantic assets development takes substantial effort. Having a proper process may help
- Having different practices of application for the same semantic asset is normal
- Having multiple semantic assets for the same domain is normal
- Semantics can be expressed and exploited using various modelling techniques and IT solutions

…but there are other flavours of interoperability beyond semantics)*

Challenge	Popular response
Syntactic interoperability	Common terminology, common XML schemas
Technical interoperability	Configurable and well-governed software, well-specified APIs
Semantic interoperability	Clear identification of all concepts, connections between them, and inference rules

)* For "layered" interpretation of these interoperability aspects, see Andreas Tolk et al. Composable M&S Web Services for Net-Centric Applications. The Journal of Defense Modeling and Simulation. Vol.3(1), pp.27-44 (2016). https://doi.org/10.1177/875647930600300104 - kindly indicated by Zachary Trautt (NIST)

… also interoperability is not the end in itself

- There is often a trade-off between interoperability and extensibility
- Use cases and success stories are important
- Tools and technology to support semantic modelling and models reuse are important – not only for IT infrastructure, but as a communication aid and as a means of discourse

FA?? -> FAIR

(not mutually exclusive) Solutions for Interoperability and Reproducibility of data-intensive R&D

- Sensible governance and quality documentation for IT implementations
- Metadata exchange format or self-documented data exchange formats
- APIs specifications (can be self-documented, too)
- OO design frameworks with well-defined objects for a specific domain
- DSLs (domain-specific programming languages)
- Schema languages / specifications, including for RDF
- Ontologies
- Workflows (for a smaller number of well-defined objects compared to the OO design approach – perhaps just one common object) and engines for the workflows execution)*

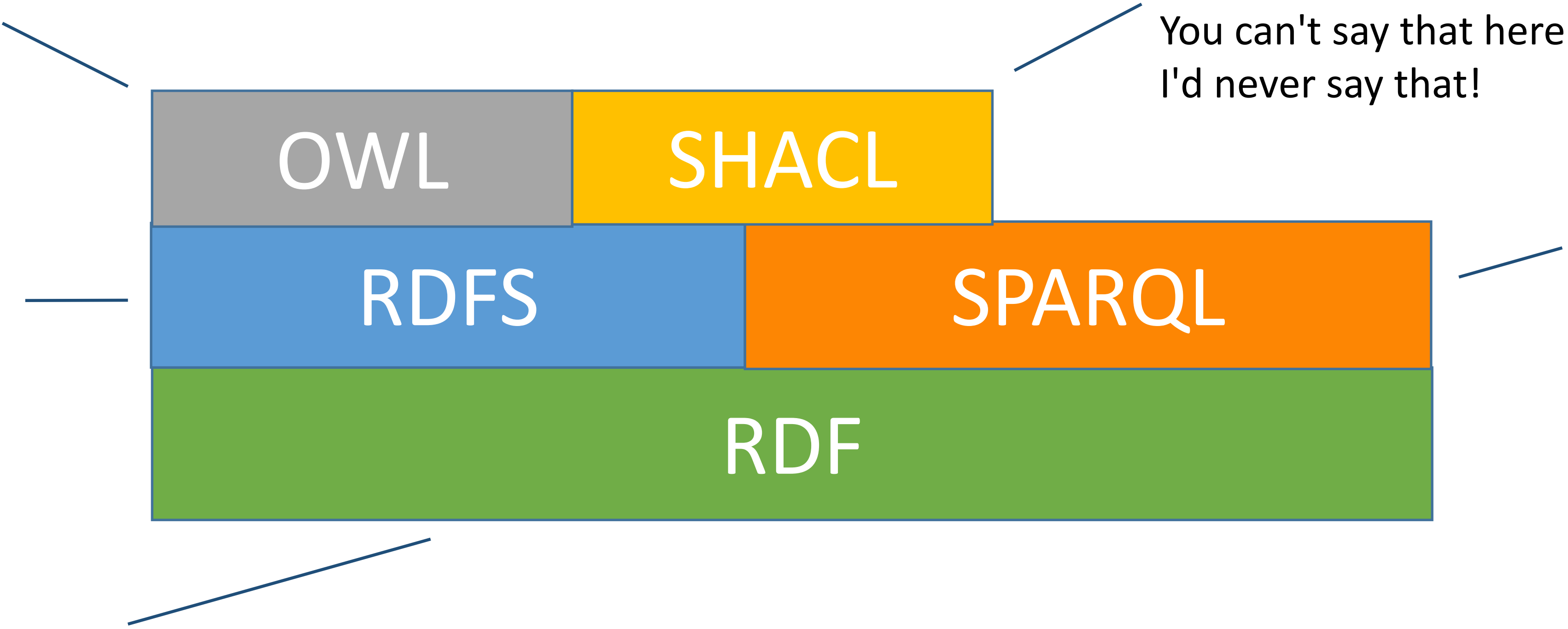
)* See Sean Bechhofer et al. "Why linked data is not enough for scientists". https://doi.org/10.1016/j.future.2011.08.004
They refer to www.myexperiment.org as a platform for the new kind of research discourse empowered by workflows

(Relatively) new kid on the block: SHACL

https://www.w3.org/TR/shacl/

What makes logical sense to say?

Is that word used correctly?
What do you need to know from me?
You can't say that here!
I'd never say that!



OWL

SHACL

RDFS

SPARQL

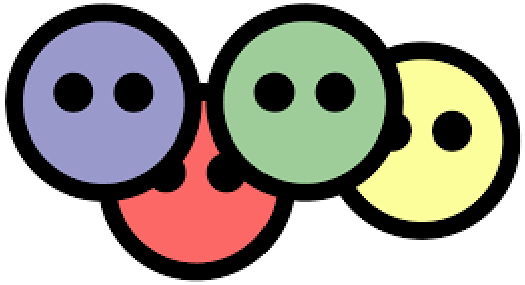
RDF

What words do we have?

What did you say about XYZ?

Statements: What is being said?

The diagram replicates the one in Richard Cyganiak’s 2016 presentation “SHACL: Shaping the Big Ball of Data Mud”
https://www.slideshare.net/cygri/shacl-shaping-the-big-ball-of-data-mud



Communication with a wider community of semantic modellers and technologists that can be beneficial for Materials Science

- Fintech / FIBO community can advise on quality governance for the ontology development. Look online, approach them directly, or I can see what I can do
- Bio-informaticians may be able to advise on management of multiple semantic assets, and on their actual use for indexing. Look online, ask EMBL-EBI (UK) – directly or using me as a proxy
- EUON (European Ontology Network) – only one workshop so far, supported by EUDAT project. If interested, ask Yann le Franc (co-chair of the RDA Vocabularies Interoperability IG) – directly or using me as a proxy
- There are pockets of European expertise in semantic modelling & visualization tools. If interested, ask Kārlis Čerāns (University of Latvia) – directly or using me as a proxy

Picture: FOAF (friend of a friend) ontology logo

Opportunities and goals for further discussions

- Semantic Assets for Materials Science task group in RDA (next call 28th November 14:00 CET)
- EMMC International Workshop in Vienna (February 2019)
- RDA groups and RDA plenary in Philadelphia (April 2019)
- DAMDID conference and a potential workshop on informatics for materials science in Kazan or Moscow (October 2019)
- Possible synergies between EMMC and Physical Sciences Data Service (with service vision developed through 2019)
- Future EU projects