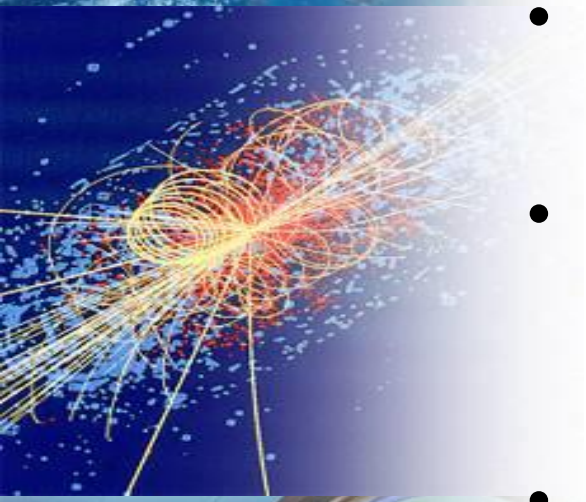
Semantic categorization of DDI metadata

EDDI12 – 4th Annual European DDI User Conference
Bergen 3-4 December 2012

Vasily Bunakov, STFC, United Kingdom
vasily.bunakov@stfc.ac.uk

engage

Science & Technology Facilities Council



Science and Technology Facilities Council

- One of Europe's largest multidisciplinary research organisations
- Operates large-scale research facilities in the United Kingdom and gives access to similar facilities world-wide
- Funds university research in physics, astronomy and space

See also: www.stfc.ac.uk

Science & Technology Facilities Council

STFC Scientific Computing Department

The StorageTek tape robot 100PB Capacity

- High performance computing including the UK's most powerful computer
- The UK hub for CERN LHC data
- Data archives:
  - ISIS: ~ 25 years, 3 mln files
  - Diamond: ~ 5 years, 100 mln files
- Data modeling, including mature metadata framework for facilities research lifecycle
- DOIs for data via DataCite / British Library
- Data access policy: promoting open access

See also: www.stfc.ac.uk/scd

Science & Technology Facilities Council

ENGAGE business case

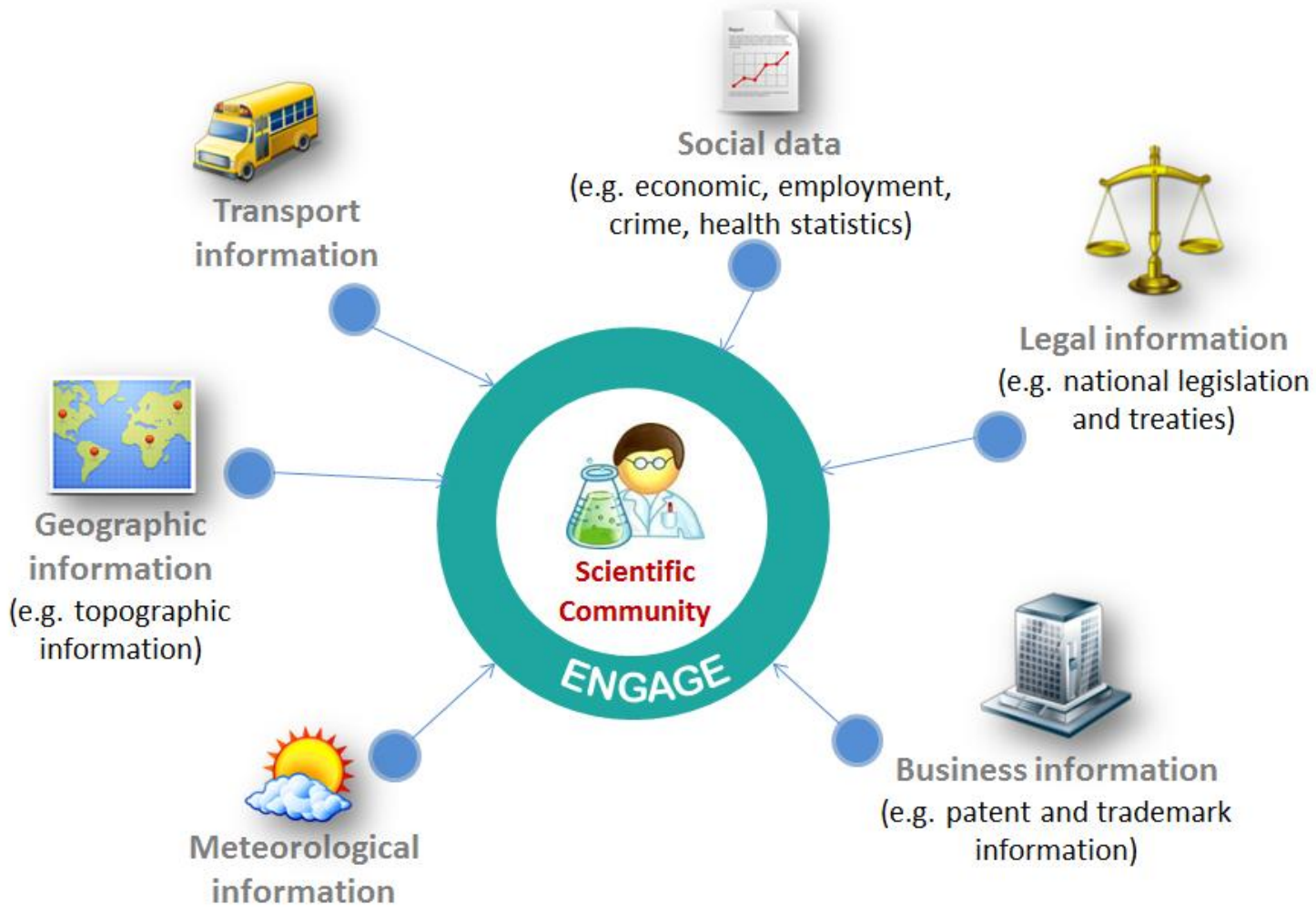
- National and local governments, as well as other public bodies are publishing lots of data on the Web
- European infrastructure is needed
- To provide Public Source Information (PSI) to research communities and citizens
- Data linking with Social Science archives is important and very welcome

engage

An Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens

www.engage-project.eu

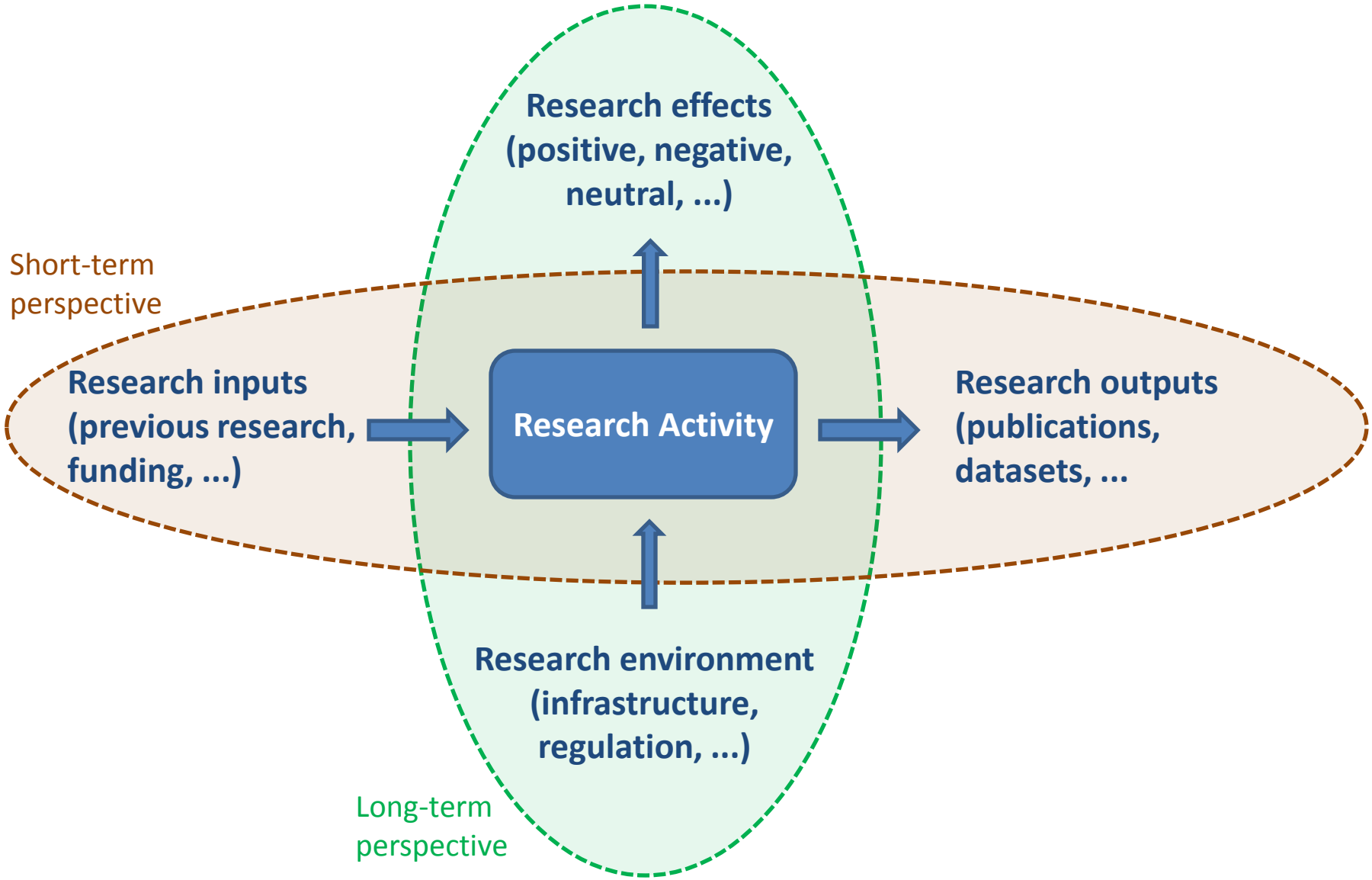
ENGAGE vision of linked data across different domains



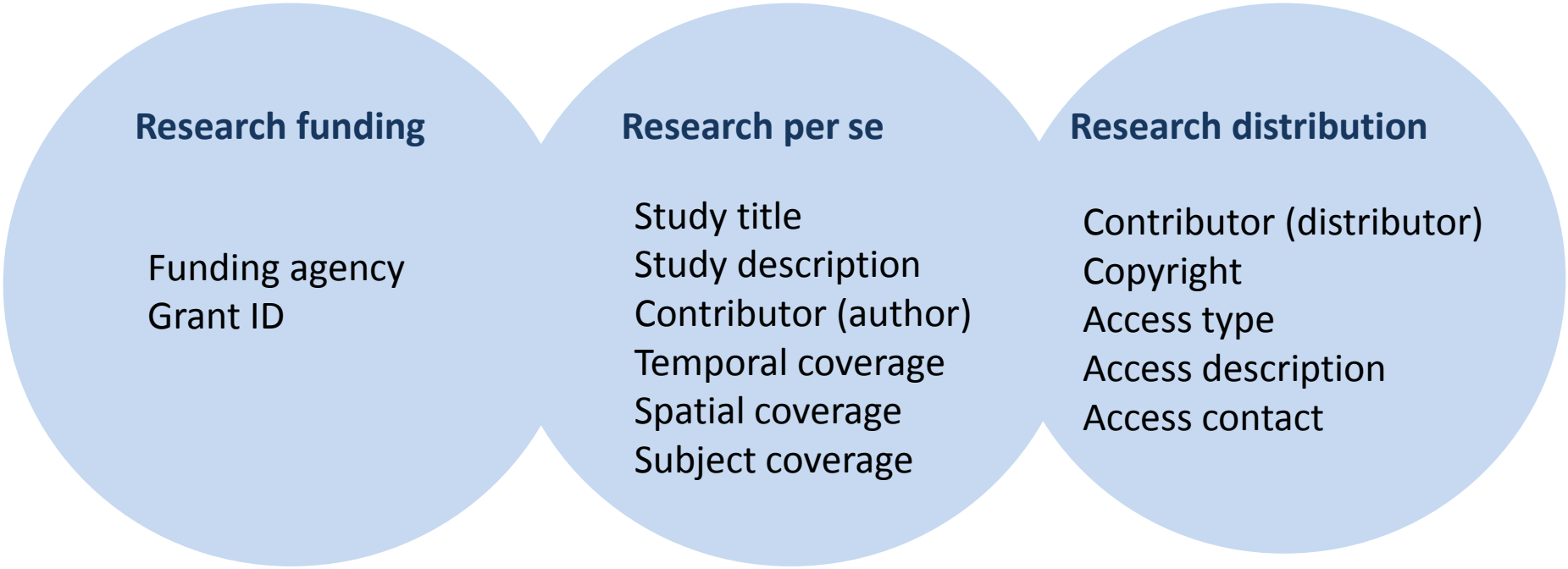
To make research data linkable, we need to reasonably model research activity

- Keep the model generic enough
- Keep it simple for better adoption and "opportunistic" application
- Aim it not at humans only but at machines / software agents, too, e.g. care about automated semantic inference

On models: our view of a research “cell” activity

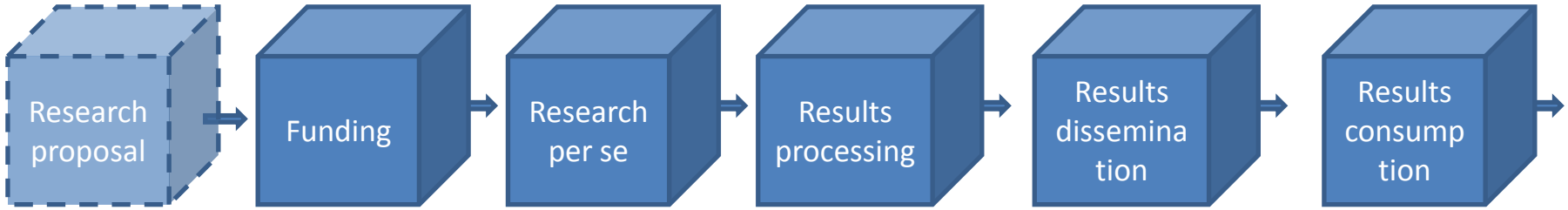


More than one research activity in one DDI record



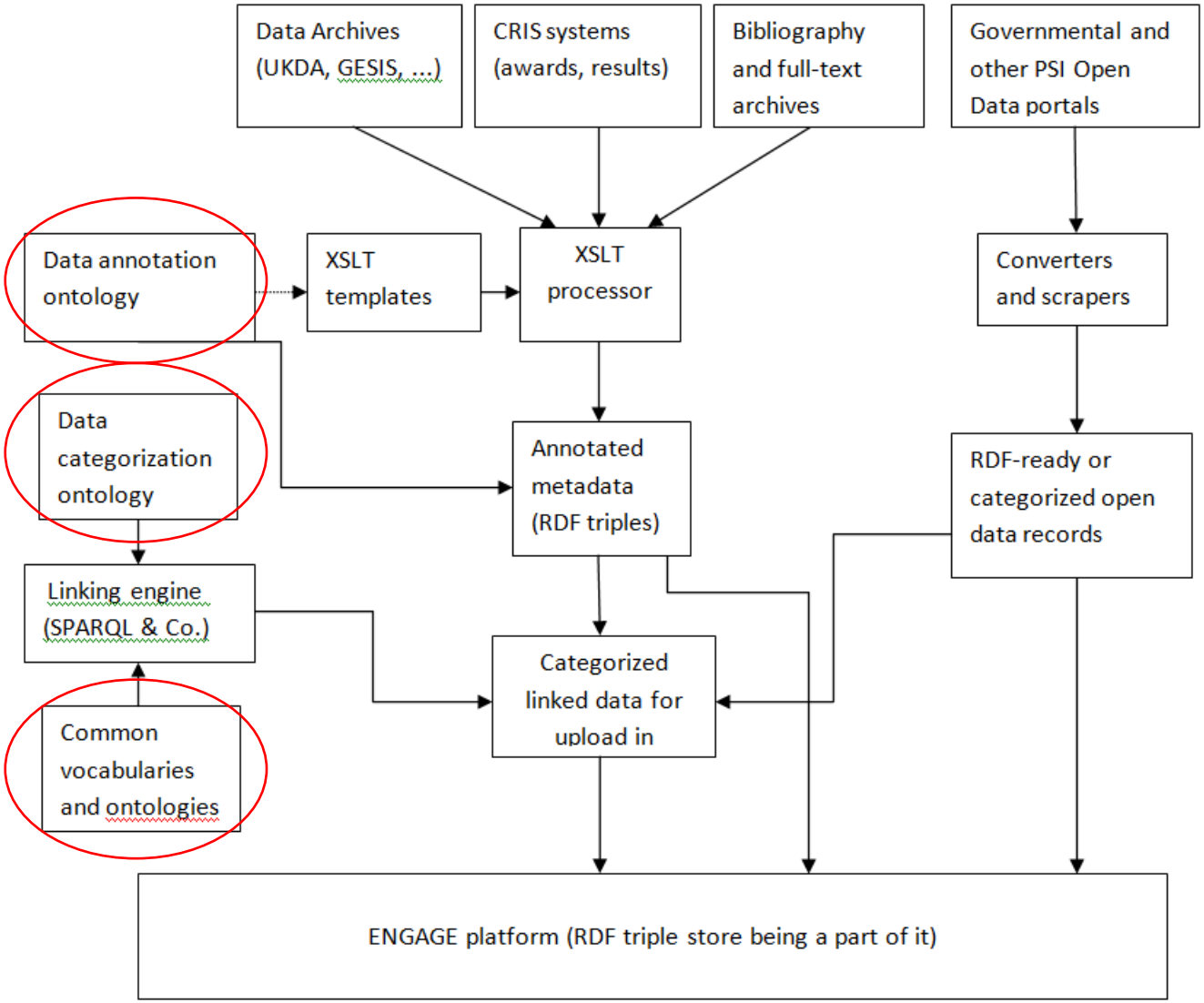
(A newer) DDI has been designed to cover the entire research lifecycle in specific branches of science but when we speak of a common infrastructure, we have to consider different information needs, and different modes of information re-use

Research decomposed in cells, with examples of entities for each



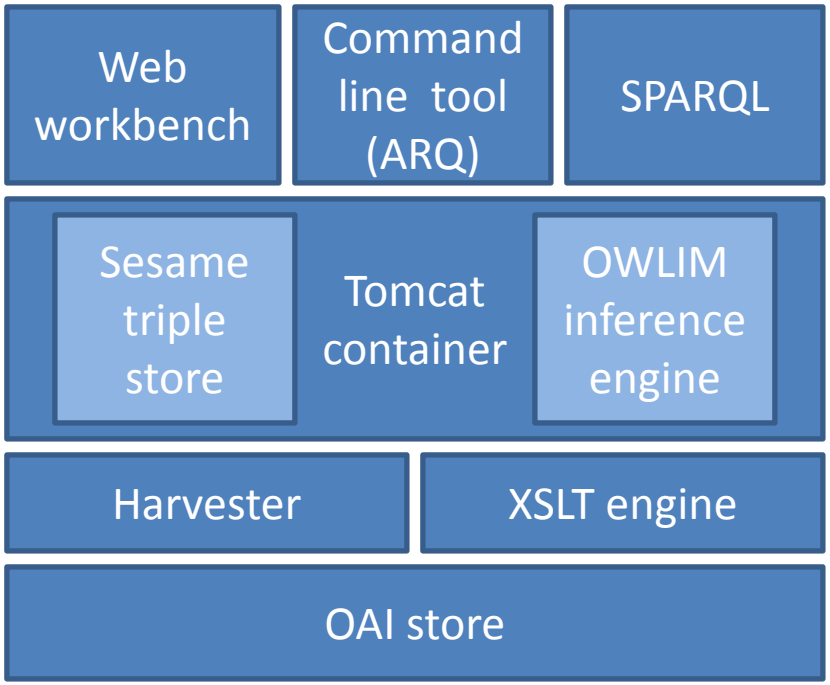
Ontology class	Funding	Research per se	Results processing	Results dissemination	Results consumption
Input	Research proposal	Award (grant)	Dataset	DDI record	DDI record or its manifestation
Output	Award (grant)	Dataset	DDI record	Web service	Feedback
Actor	Researcher candidate	Contributor (author)	Data archive	Dissemination service	Web service user
Effect	Researcher's department budget	Whatever is claimed in proposal	Economical effect of processing	Economical effect of distribution	Impact on further research
Condition	Funding body rules & regs	Microdata regulation	Data processing guidelines	Data access regulation	Research purpose statement
Scope	Certain branch of science	Certain geolocations	National research	International research	Certain HASSET keywords

Ontologies in the context of ENGAGE data processing



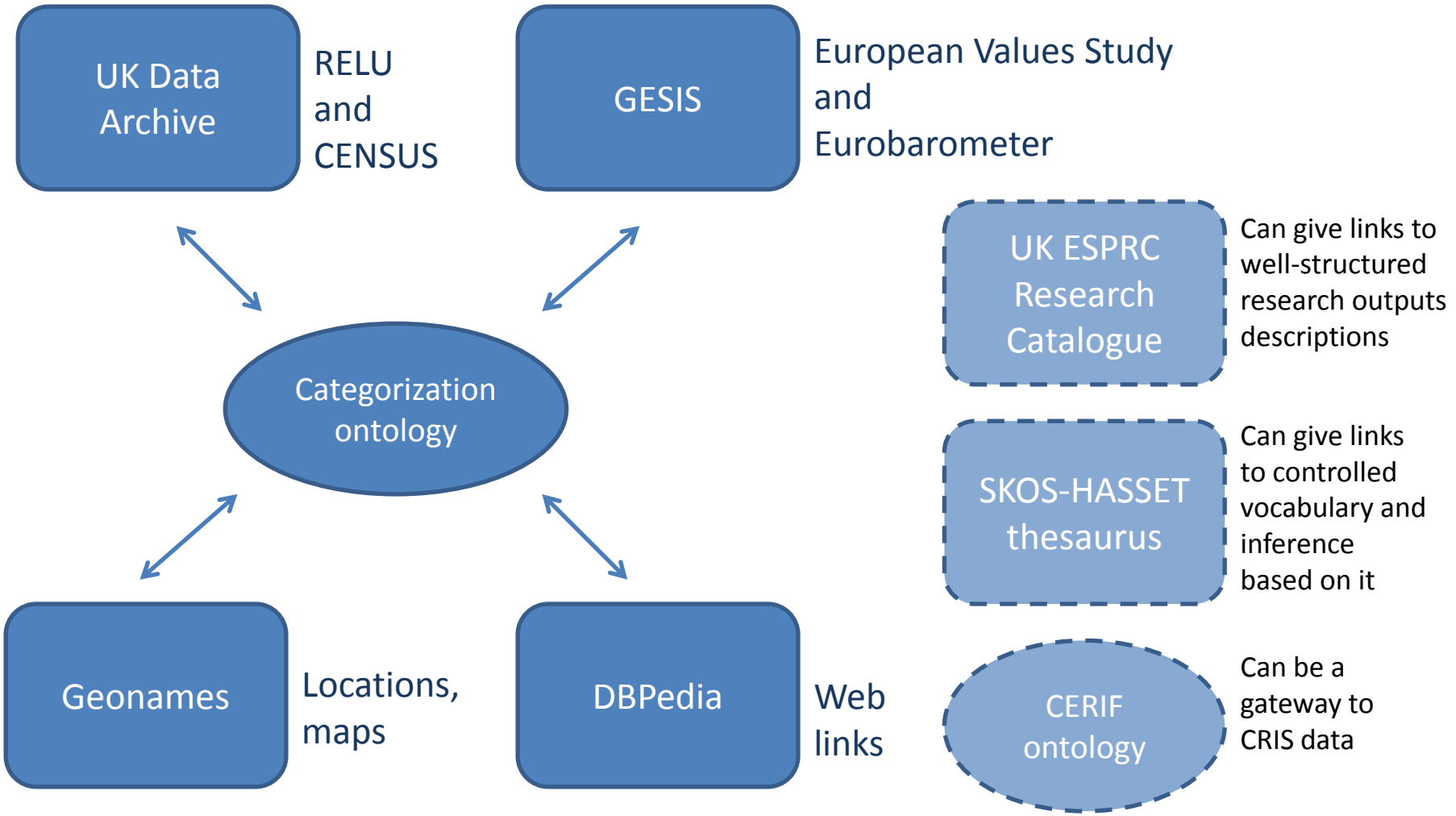
Technology stack for experiments with DDI metadata

- OAI protocol and harvesters
- Tomcat, Sesame, OWLIM, ARQ
- XSLT, SPARQL



Just enough to support “RDFS Plus” data modeling

Data sources and models tried out and planned



“Shallow” categorizations possible against DDI metadata

- **Inputs and Outputs**
  Defined via Categorization Ontology to facilitate data discovery and data provenance tracking
- **Geolocations**
  For research scope, for actors, ...
- **Subjects**
  At least, their types: HASSET, ZA-Categories, ...
- **Actors**
  At least, their types: investigators, distributors...

Any deeper categorizations will require Subject Matter Experts powered by data refinement & linking tools (Google Refine, LOD2 Silk, ...)

Why we need semantics?

Location	No. of references to Location	Part Of (parent)	Same As
"UNITED KINGDOM"	16		"GB United Kingdom"
"ENGLAND"	14	"GREAT BRITAIN"	
"SCOTLAND"	13	"GREAT BRITAIN"	
"WALES"	10	"ENGLAND AND WALES"	
"GREAT BRITAIN"	8	"UNITED KINGDOM"	"GB-GBN Great Britain"
"ENGLAND AND WALES"	6	"GREAT BRITAIN"	
"GB-GBN Great Britain"	5		"GREAT BRITAIN"
"GB-NIR Northern Ireland"	5		"NORTHERN IRELAND"
"NORTHERN IRELAND"	5	"GREAT BRITAIN"	"GB-NIR Northern Ireland"
"PEAK DISTRICT"	4	"ENGLAND"	
"GB United Kingdom"	3		"UNITED KINGDOM"
"SOUTH WEST ENGLAND (REGION)"	3	"ENGLAND"	
"EAST MIDLANDS (REGION)"	1	"MIDLANDS"	
"MIDLANDS"	1	"ENGLAND"	

UK Data Archive
GESIS

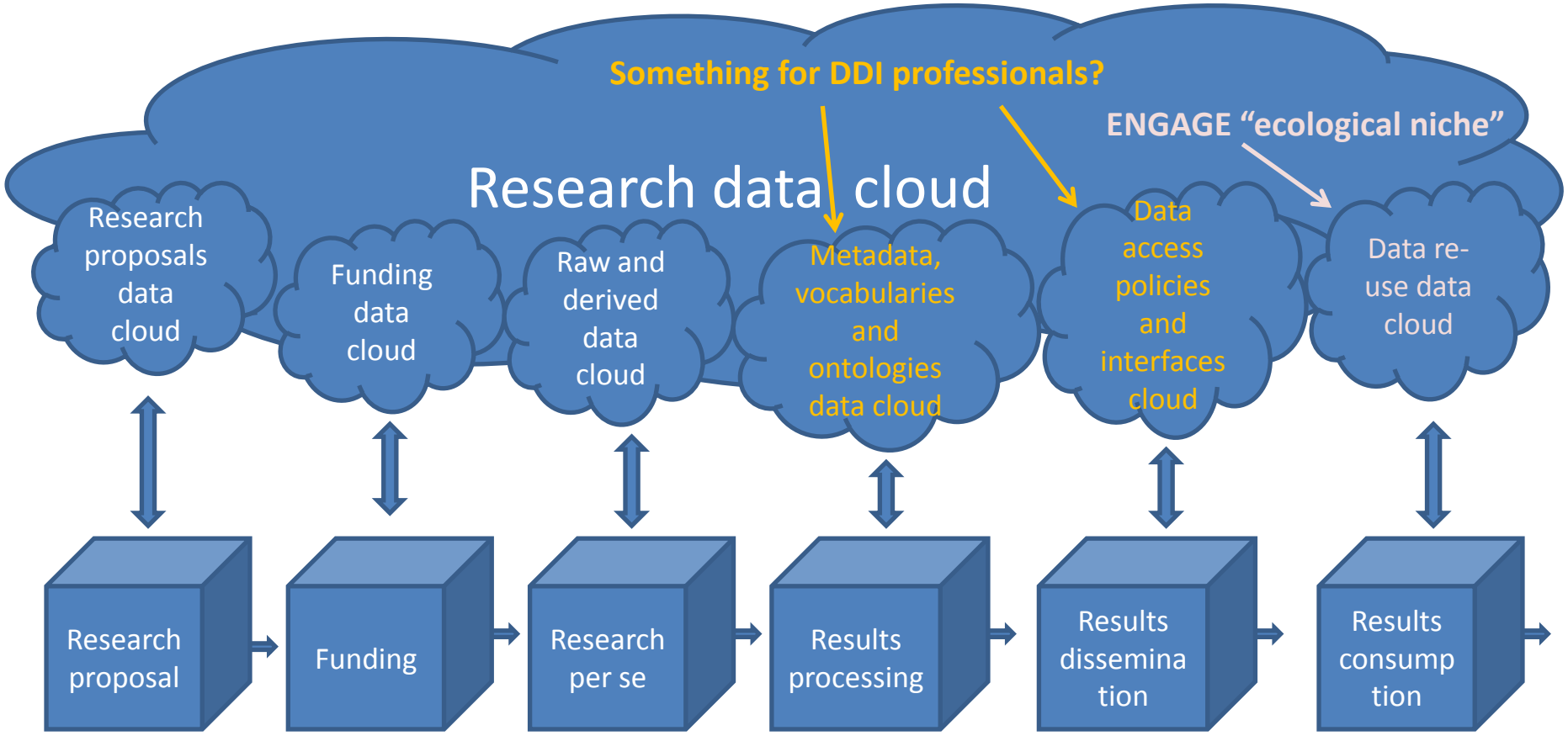
Good news about Linked Data

Easy to create new entities...	CONSTRUCT {?study engage:hasScope ?location . ?location a ukda:Location . ?location engage:name ?name . } WHERE {?study a ukda:Study . ?study ukda:locationKeyword ?name . BIND (URI(CONCAT("http://example.org/stuff/engage#",str(?study),ENCODE_FOR_URI(?name))) AS ?location) }
...and generalize them...	ukda:Location rdfs:subClassOf engage:Location . gesis:Location rdfs:subClassOf engage:Location . select ?study where {?study engage:hasScope ?location . ?location a engage:Location . }
...as well as properties...	ukda:locationKeyword rdfs:subPropertyOf engage:locationKeyword . gesis:locationKeyword rdfs:subPropertyOf engage:locationKeyword . select ?study where {?study engage:locationKeyword “Nothern Europe" . }
...also make “sameness” claims	gesis: GB-NIR_Northern_Ireland owl:sameAs ukda:NOTHERN_IRELAND .
Then it is fairly easy to link data to the “cloud”	gesis:EurobarometerSeries owl:sameAs dbpedia:<http://dbpedia.org/resource/Eurobarometer> . ukda:University_of_Essex rdfs:seeAlso geonames:<http://sws.geonames.org/6690170> .

Not so good news (challenges)

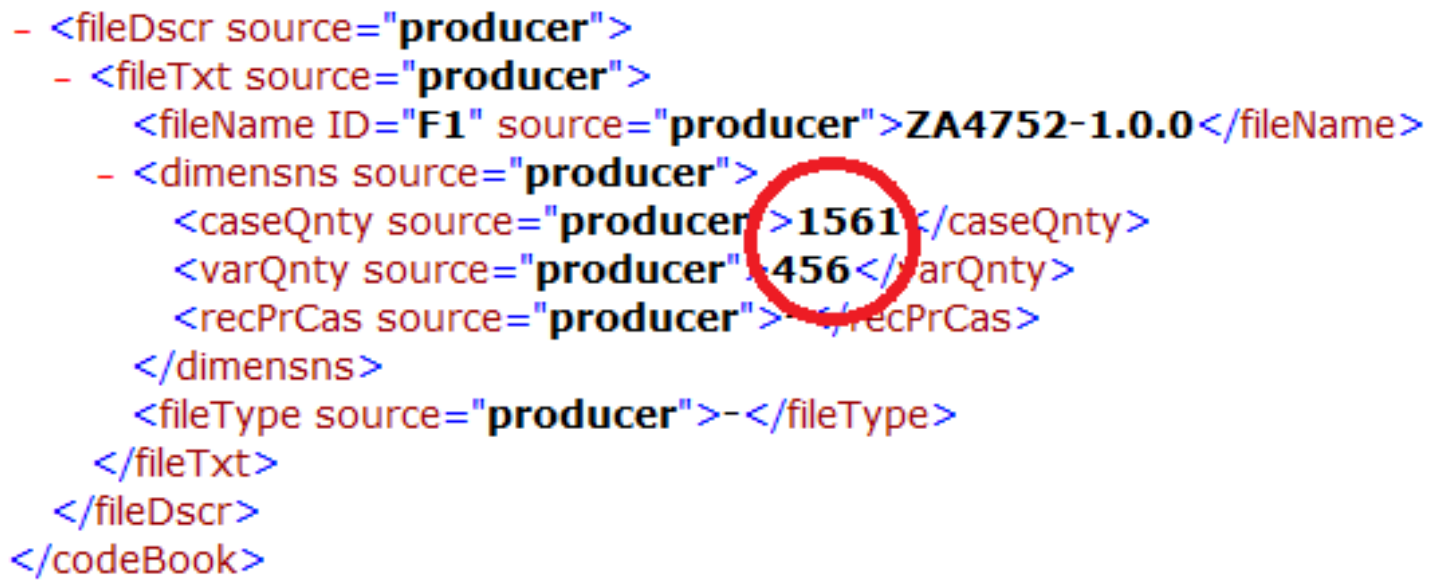
- Someone has to contribute to Linked Data cloud before you link to it, or re-use parts of it
- That someone may not have enough incentives for contribution
- Intellectual property, copyright and regulation can be “natural” limitations
- Data practitioners need a proper discussion on the above, as well as their best practices shared

Research data as “cloud of clouds”



See also: [Tim Berners-Lee on “bag of chips”]

A challenge of Linked Data processing on a granular level



A) One European Values Study dataset may result in hundreds of thousands RDF triples

B) “A rule of thumb” is that an average triple store instance can handle 1,000,000,000 triples

A + B => granular Linked Data processing of just a dozen European Values Study datasets may require a dedicated triple store

Open questions and suggestions

- Use cases for linking DDI (meta)data with .gov and other PSI (meta)data are very welcome
- Insufficient openness of DDI data sources may hinder Linked Data developments for them
- Linked Data on granular level is Big Data so we in natural sciences are out there with a computer power

engage

www.engage-project.eu

Science & Technology Facilities Council

www.stfc.ac.uk

Thank you!


data
research
LHC
e-Science
used
particle
petabytes
Tier-1
collisions
experiments
energy
STFC
first
November
physics
services
researchers
construction
experimental
rates
computing
users
year
new
R89
sites
UK
Rutherford
Appleton
Laboratory
processing
CERN
machine
events
CMS
beams
network
centre
world
Atlas
expected
storing
started
project
number
years
recorded
science
programme
also
simulating
requirements
time
initiatives
support
economic
distributed
experiment
activities
plasma
smoothly
room
protons
handling
including
wide
high
Two
TeV
March
impact
universe
store
four
tier-1
start
move
last
analysis
received
main
Research
LHCb
SCARF