CDISC

2018 EUROPE INTERCHANGE BERLIN 23-27 APRIL

2018
EUROPE INTERCHANGE

CDISC

Draft publication – release q3 2018

Title: Implementing data standards for effective data sharing and essential steps towards optimising medical research.

Authors: Paul Houston[1], Larry Callahan[2], Michael Braxenthaler[4,] Lauren Becnel[1], Sam Hume[1], Dorina Bratfalean[1], Martin Romaker[3], Philipe Roca-Serra[3,] Andreas Tilman, Susanna Sansone, Ibrahim Emam, Kerstin Forsberg, Frederik Malfait, John Ezzell, Frank Petavy

Affiliations:

1. CDISC and CDISC European Foundation
2. FDA –Office of Health Informatics, Silver Spring Md USA.
3. Oxford Research University –
4. Roche
5. Biosci Consulting
6. astrazeneca

Etriks IMI Grant agreement number	115446

CDISC

Effective data sharing definitions

- 'Effective data sharing' –relies upon 'high quality structured data' being fully shared, without limitations, for open interrogation and for aggregation with complementary data sets.
- 'Effective data' or 'High quality structured data' – Implementation of a rigorous data standards environment following approaches such as FAIR (Findable, Accessible, Interoperable and Reuseable) and The Dublin Core Interoperability levels.
- Semantic Interoperability – when high quality data is inter-connected on the internet in a meaningful way to create new knowledge and medical breakthroughs

Recent history and the data sharing landscape

- Enshrined in European Public Policy – ‘open access to document’ Vienna agreement
- Cochrane Institute and Ben Goldacre consistent voices
- Reaction is EMA policy 0070 – ‘Study reports’ ‘summary data and results’ next steps will be Patient Level Data
- Pharma response – Datasphere.com ClinicalStudyDataRequest.com , academic Yoda, AllTrials, Vivali
- European Commission making research data open by default
- WHO, Wellcome, Médecins Sans Frontières etc statement

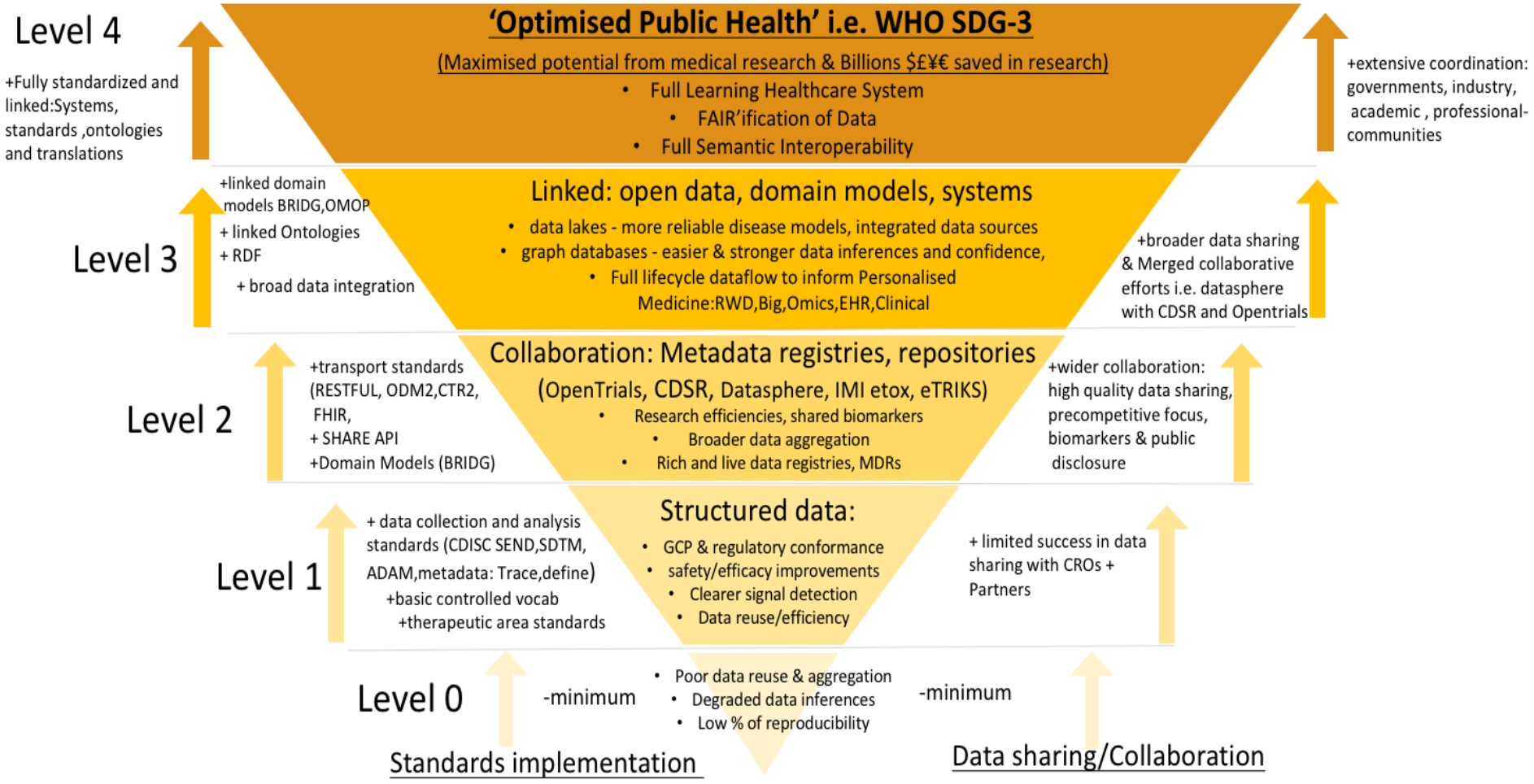
with good structured data we are creating knowledge
aggregating knowledge into wisdom is the aim
wisdom that leads to medical breakthroughs our goal

Motivations/Why?

- Connect the worlds healthcare data (costs per year of maintain and achieve of $300-700 million per year for semantic interoperability but potential savings in healthcare of $30 billion per year in the USA alone
- McKinsey have made other estimates for cost savings for open data in healthcare of up to $300 billion.

- Why – to find cures, more precise medicines, improved public Health, save lives, create efficiencies

2018
EUROPE INTERCHANGE



5 top problems

1. Data standards are not consistently mandated/recommended by govt/funders/foundations
2. Standards are not applied consistently, implementation curves too steep, current low MDR use,
3. Data isn't shared fully: fenestrated data or licencing or anonymization issues
4. More resources: Creating and maintaining semantic interoperability and standards is complex the sector needs more hands, more funding, more government support.500,000 data stewards.
5. More co-ordination less innovation - and less duplication of terminologies and Domain Models.

1. Creating data atoms quality high enough to be Interoperable and Re-useable/reproducible , the essential I & R of the FAIR data principles.
2. For data sharing essential to used established anonymization techniques i.e. Phuse – Feran, Emman et al 2015
3. It is the clinical research findings that are shared which are the foundation to stratifying medicines for precision medicine
4. Controlled vocabularies and shared ontologies, Biomedical concepts, CDISC RDF and RDFS and SKOS for the semantic Web.

Use Case: Improve registries via structured and connected data: Clinical research & Patient/Disease

Avoiding Vioxx incident - Had the data for Vioxx and other similar cox2 inhibitor drugs such as Naproxen been made available many of the estimated 40,000 deaths might have been avoided and Merck would have avoided the 4.85 billion dollar joint law suit.

- Reporting of clinical trials is now mandatory in USA and Europe but that reporting must be timely – late submissions ignored for over 1billion in fines
- Standards exist for ingredients of drugs to be recorded in a structured format – FDA working on IPD dictionary – so medicines using like ingredients can be cross analysed
- Results for Preclinical studies are not mandated by law in Europe whereas SEND is mandated in USA. pre-clinical data is an essential ethical right Anderson & Kimmelman 2012
    - Early detection of ADRs i.e. BIA 10-2474 preclinical studies showed deaths in primates at high dosage, same in first in man studies – 1 death and several serious ADRs.

- C. Glenn Begley et al identified 53 preclinical ‘landmark’ oncology papers in top journals. Only 47 of 53 could be replicated

- ‘The loss of empirical studies are sinkholes in the medical landscape’
- Grave danger in clinical interventions being made on poor data assumptions
- With the future being AI and deeplearning data must be solid

Collaboration has already facilitated:

- Broad consensus on Domain Models : – BRIDG, OMOP, CIMI, IDMP
- Implement Meta data registries for share consistent standards implementation SHARE
- Data Sharing efforts OpenTrials, CDSR,Datasphere
- Precompetitive efforts such as IMI showing success

Deeper collaboration needed…

- Complete beginning to end standards: extending protocol, trial registration and results summaries
- Convergence of standards to enrich the data i.e. CDISC with IDMP, ODM2.0/FHIR –
- Further sharing of pre-competitive data for de-duplication of effort i.e. Drug repurposing
- Elimination of placebo arms?

CTR2 Case Study – https://www.cdisc.org/ctr2-project

- Registry information is key to informing decision making:better registries means better business analytics for pharma and future A.I initiatives,
- better pharmacovigilance,
- De-duplication of effort i.e. more effective drug repurposing
- A CTR2 standard would better inform future EDC templates
- Efficiency and cost savings in disclosure of trials

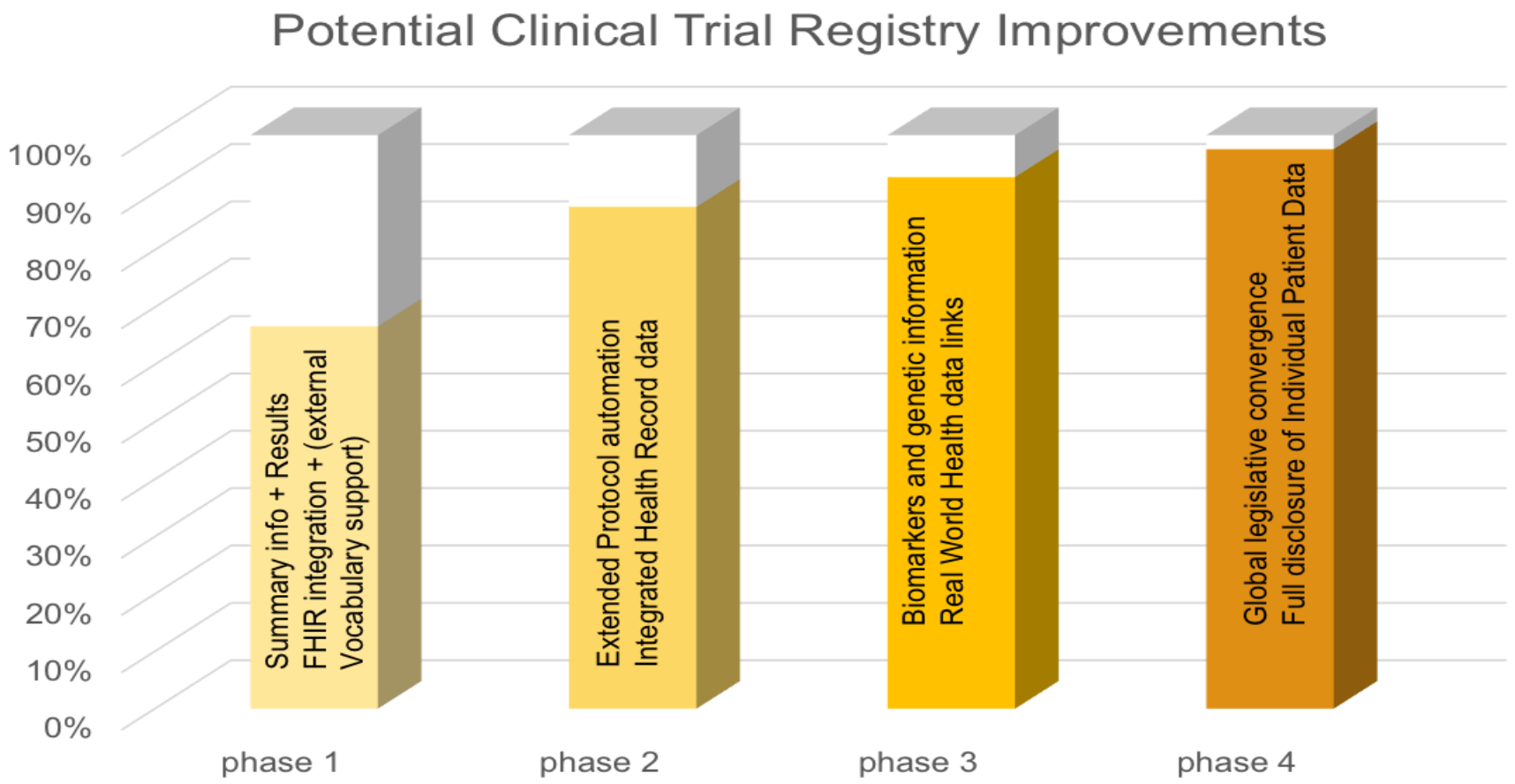


Figure 3. Moving registries towards fully structured registries with broad integrated data sources inkeeping with the 'hierarchy of data returns' paradigm

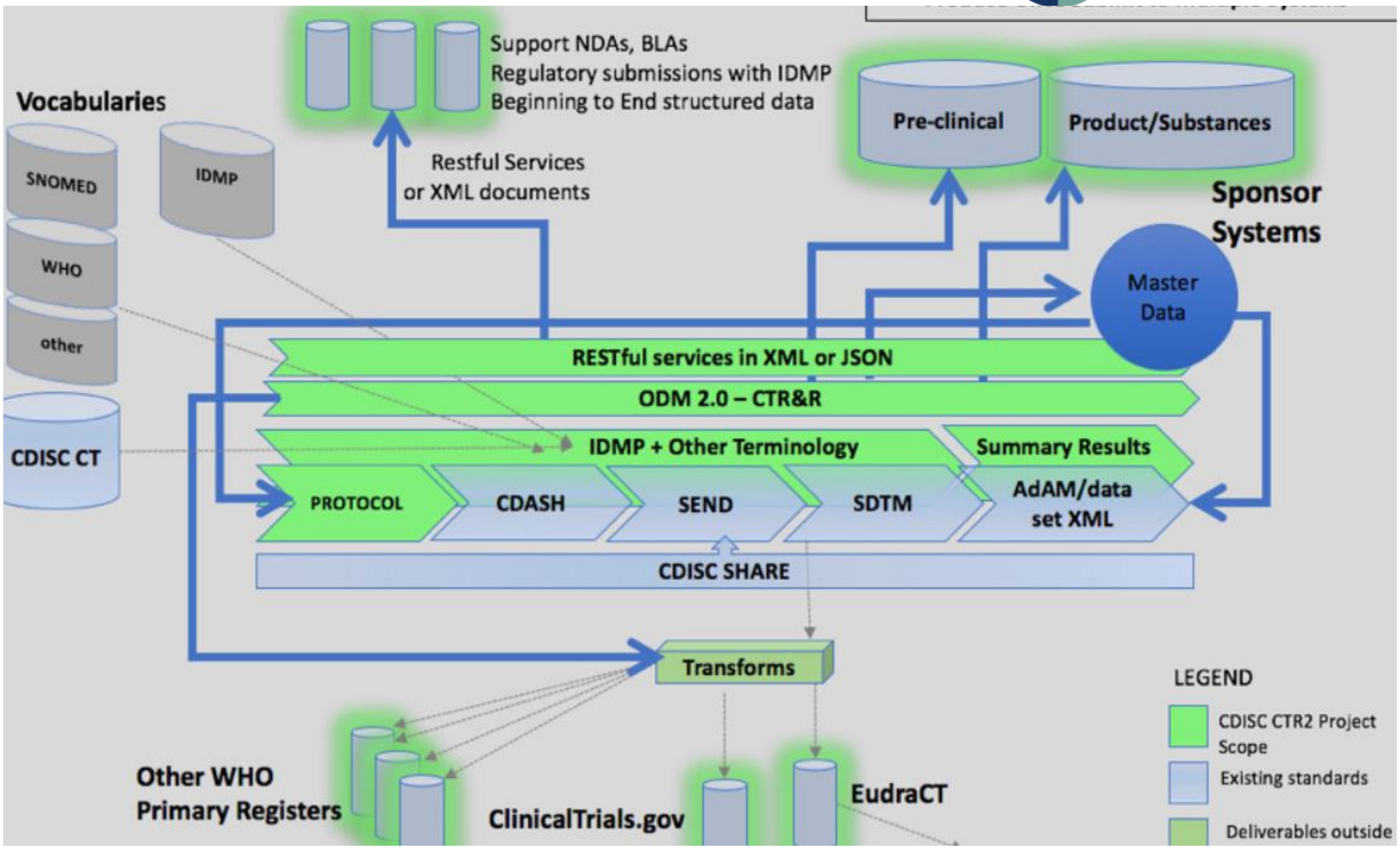


Figure 3.1 Moving registries towards fully structured registries with broad integrated data sources inkeeping with the ‘hierarchy of data returns’ paradigm

Level 3

+linked domain models BRIDG,OMOP
+ linked Ontologies
+ RDF
+ broad data integration

Linked: open data, domain models, systems

- data lakes - more reliable disease models, integrated data sources
- graph databases - easier & stronger data inferences and confidence,
- Full lifecycle dataflow to inform Personalised Medicine:RWD,Big,Omics,EHR,Clinical

+broader data sharing & Merged collaborative efforts i.e. datasphere with CDSR and Opentrials

Cypher: MATCH (drug:Drug)<-[r1:DRUG_TARGET]-(p1:Protein)-[r2:SEQ_SIM]-(p2:Protein)-[r3:BIOMARKER]-({DiseaseName:"ASTHMA"}) RETURN r1, r2, r3 limit 3(

Legend:
Nodes – concepts: Asthma: red; proteins: blue; drugs: green.
Edges – relationships among concepts: protein sequence similarity, protein-disease biomarker, protein-drug-target.

"life sciences involves modeling of an incomplete and ever-changing model of how our bodies work and what we know about it."

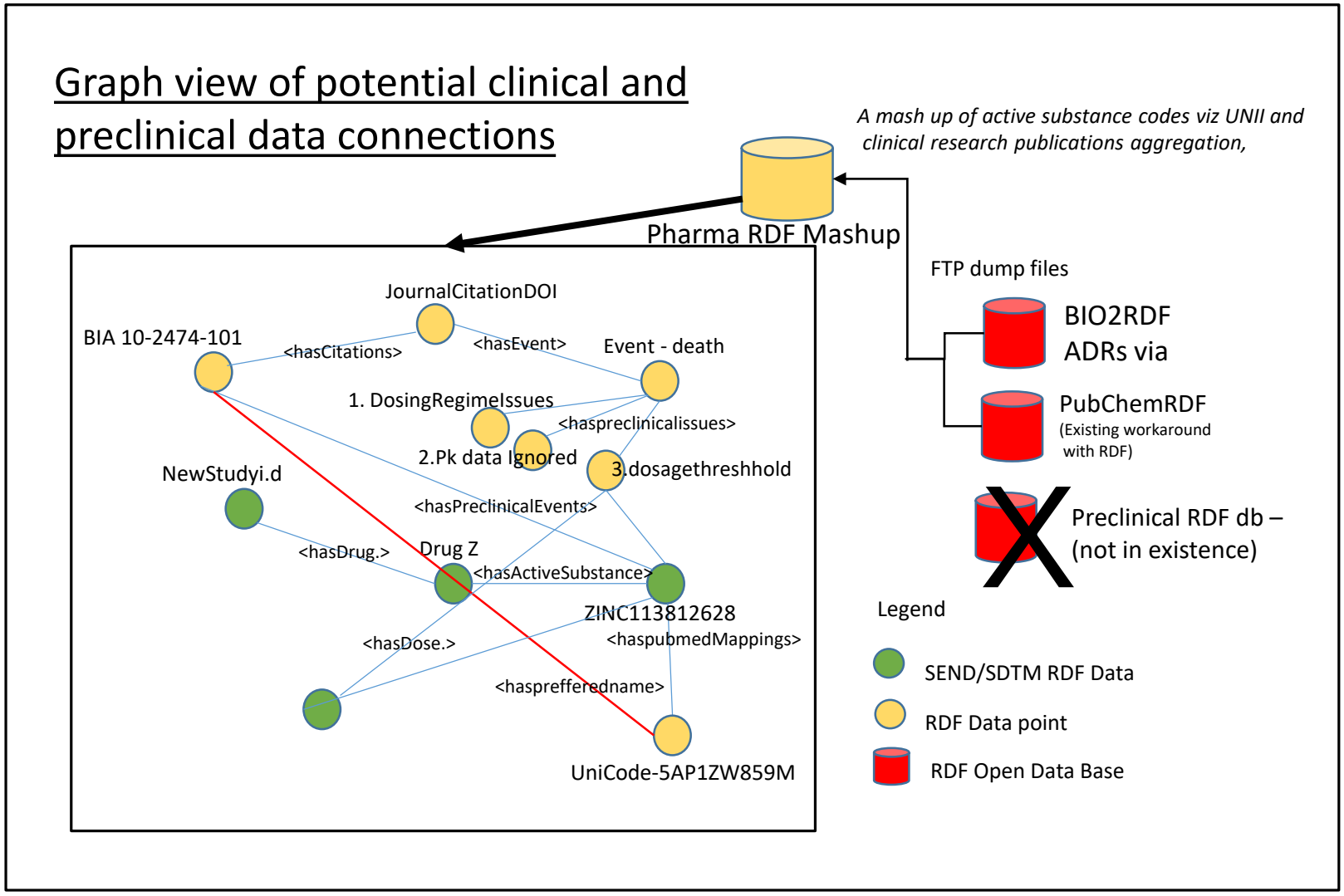


Figure 4.1 'Effective Data Sharing' Houston , Callahan et al , GRAPH USE CASE FOR BIA 10-2474-101 fatty acid amide hydrolase inhibitor

RDF Triples GRAPH USE CASE FOR BIA 10-2474-101

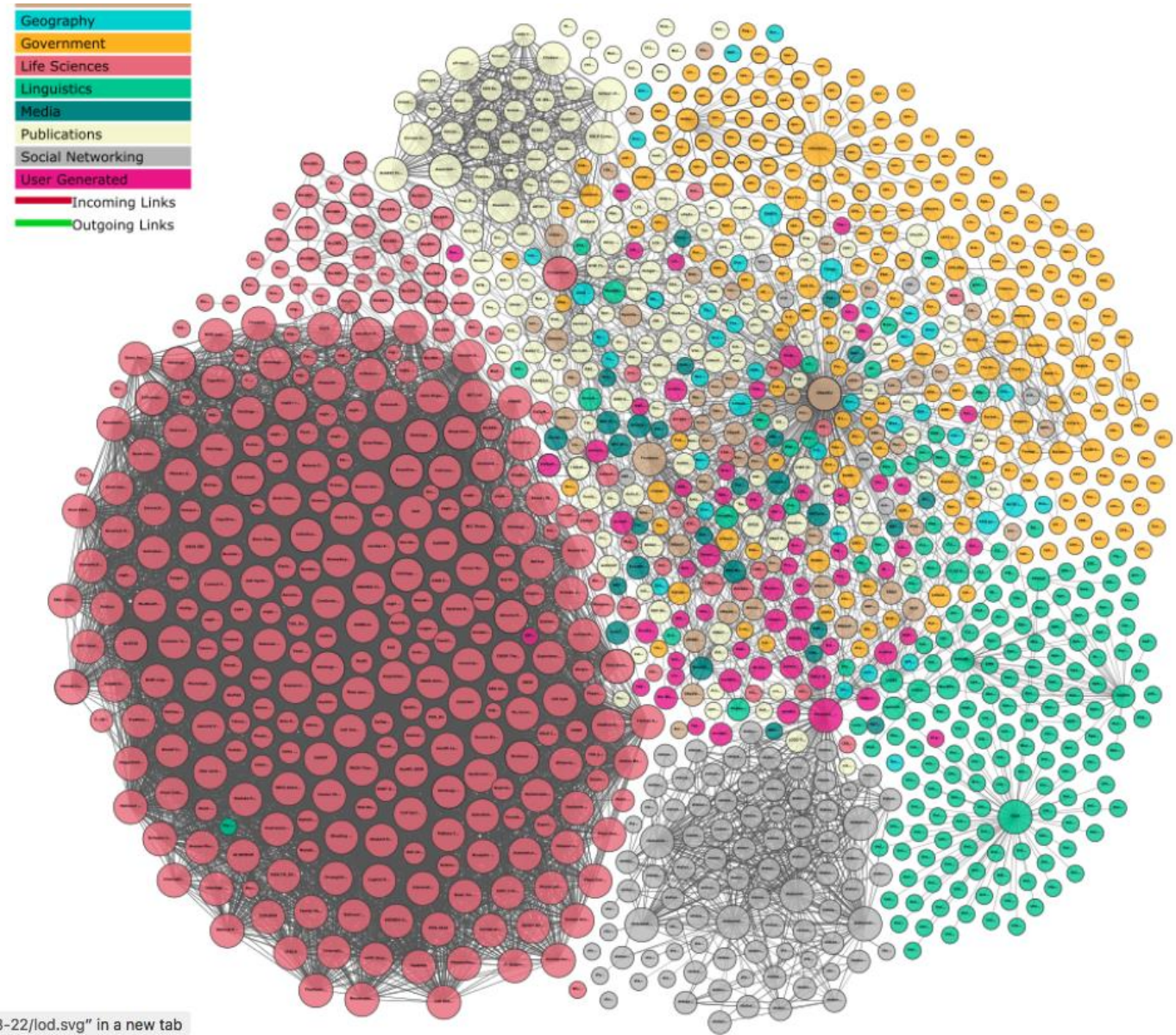
You can keep the primary sources of data in standard relational or XML-database format, but export key "facts" as triples in RDF.

https://www.w3.org/2001/sw/sweo/public/UseCases/Pfizer/

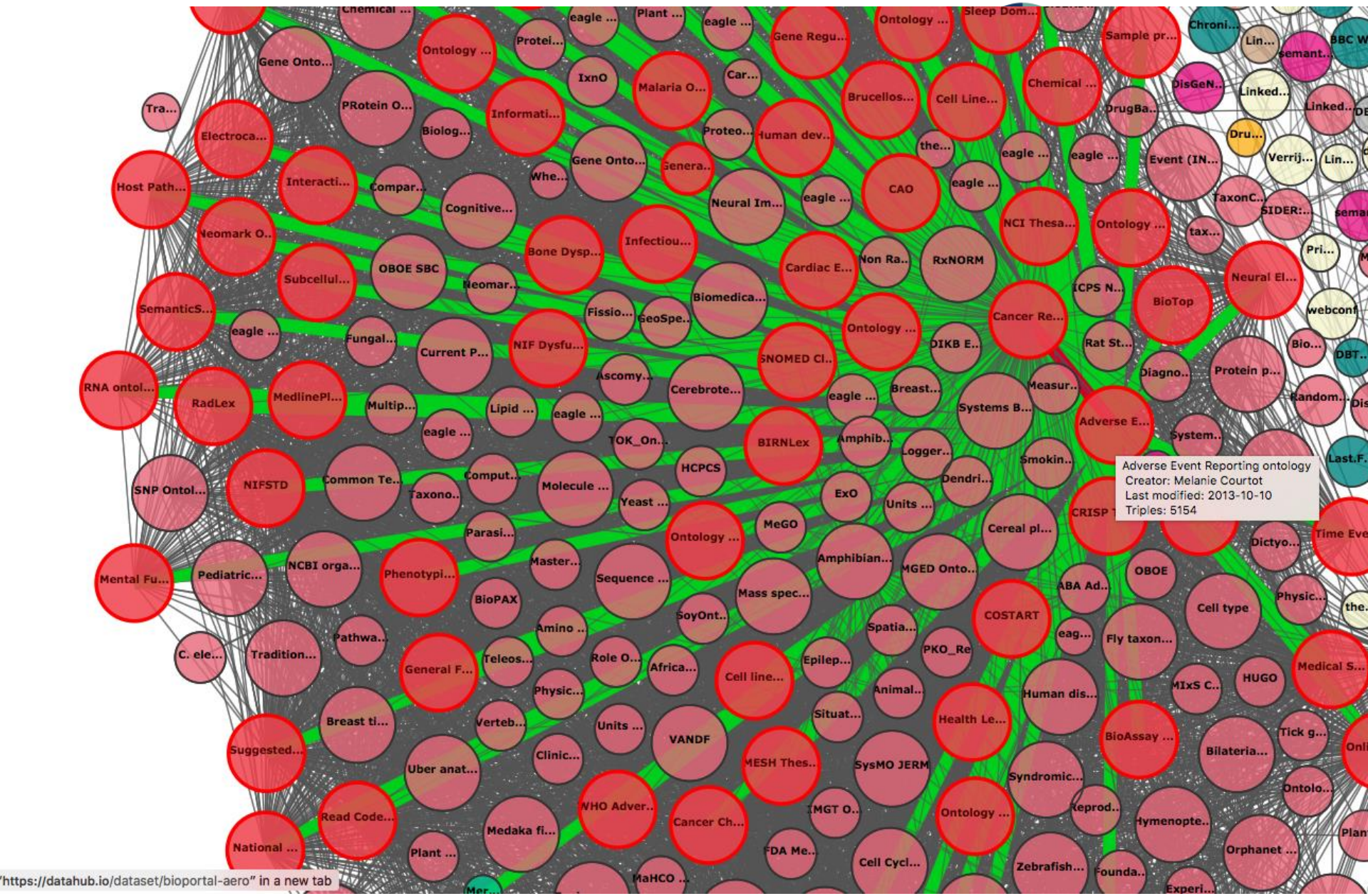
ID	Subject	Predicate	object
1	Drug Z	<hasActiveSubstance>	ZINC113812628
2	ZINC113812628	<hasMappings>	UNII-5AP1ZW859M
2	UNII-5AP1ZW859M	<HasPrefferedName>	BIA10-2474-101
4	BIA10-2474-101	<hascitations>	Journal DOI
5	Journal DOI	<hasEvents>	Event:Death of Monkeys
6	DeathofMonkeys	<haspreclinicalissues>	DosageThreshold
7	DeathofMonkeys	<haspreclinicalissues>	PK data ignored

CDISC

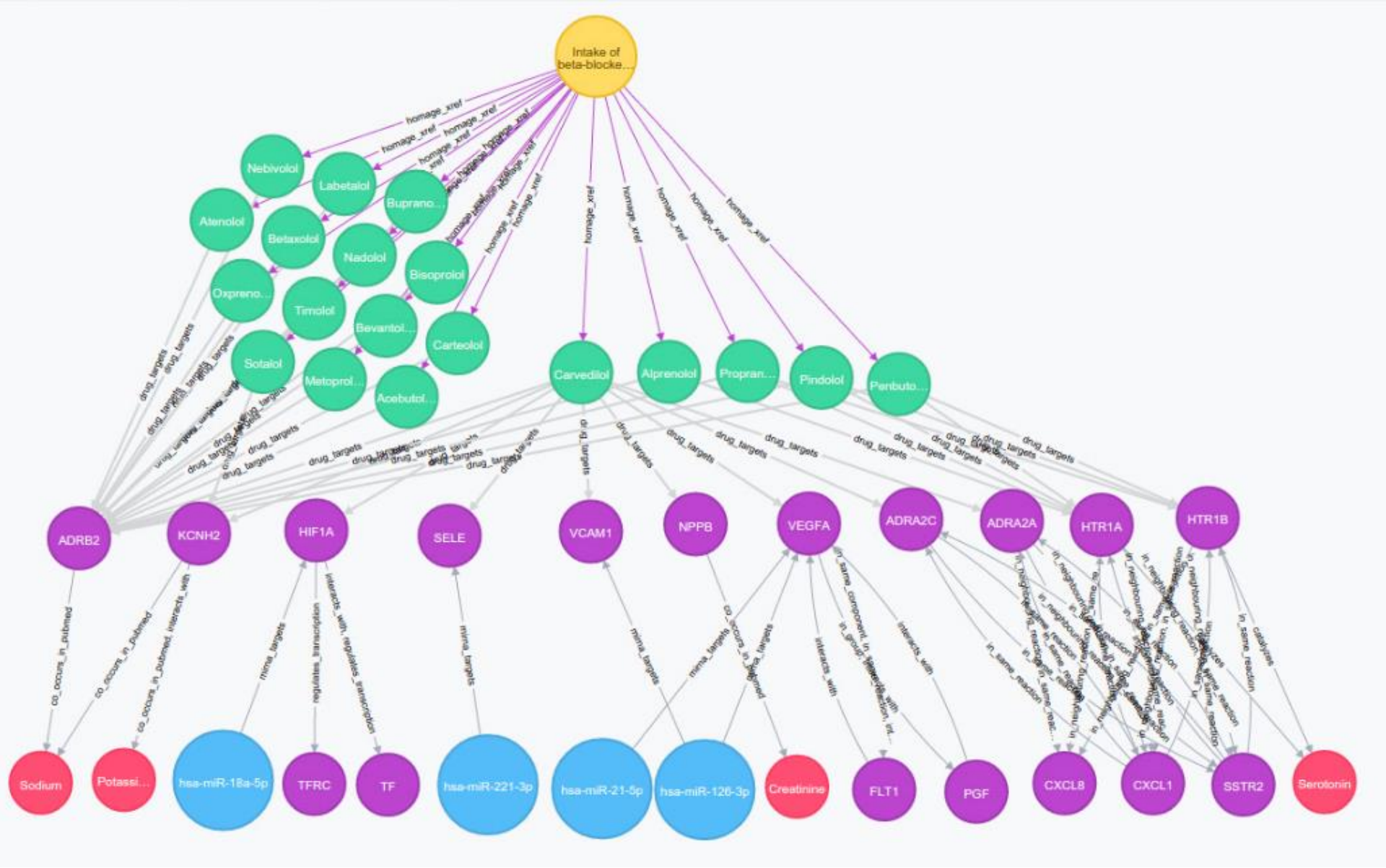
Shared Linked Data Cloud -



08-22/lod.svg" in a new tab



by September 2011 there were 31 billion RDF statements , 504 million RDF links



Today, we even have [medical doctors playing with Cypher queries](), demonstrating another important implication of this project: a shift in mindset towards truly interdisciplinary effort of clinicians and data scientists.

CDISC

RDF potential

- Once expressed in RDF, information can be represented, accessed, computed, integrated, and exchanged without the need for any translations
- provides a universal, mathematically precise, and computable language that can express a wide range of information – ideal for integrating wide data sources
- platform independence and semantic interoperability are inherent

The CDISC Mission and Principles :

- Recognize the ultimate goal of creating regulatory submissions that allow for flexibility in scientific content and are easily interpreted, understood, and navigated by regulatory reviewers.
- Acknowledge that the data content, structure, and quality of the standard data models are of paramount importance, independent of implementation strategy and platform.
- Work with other professional groups to encourage that there is maximum sharing of information and minimum duplication of efforts.

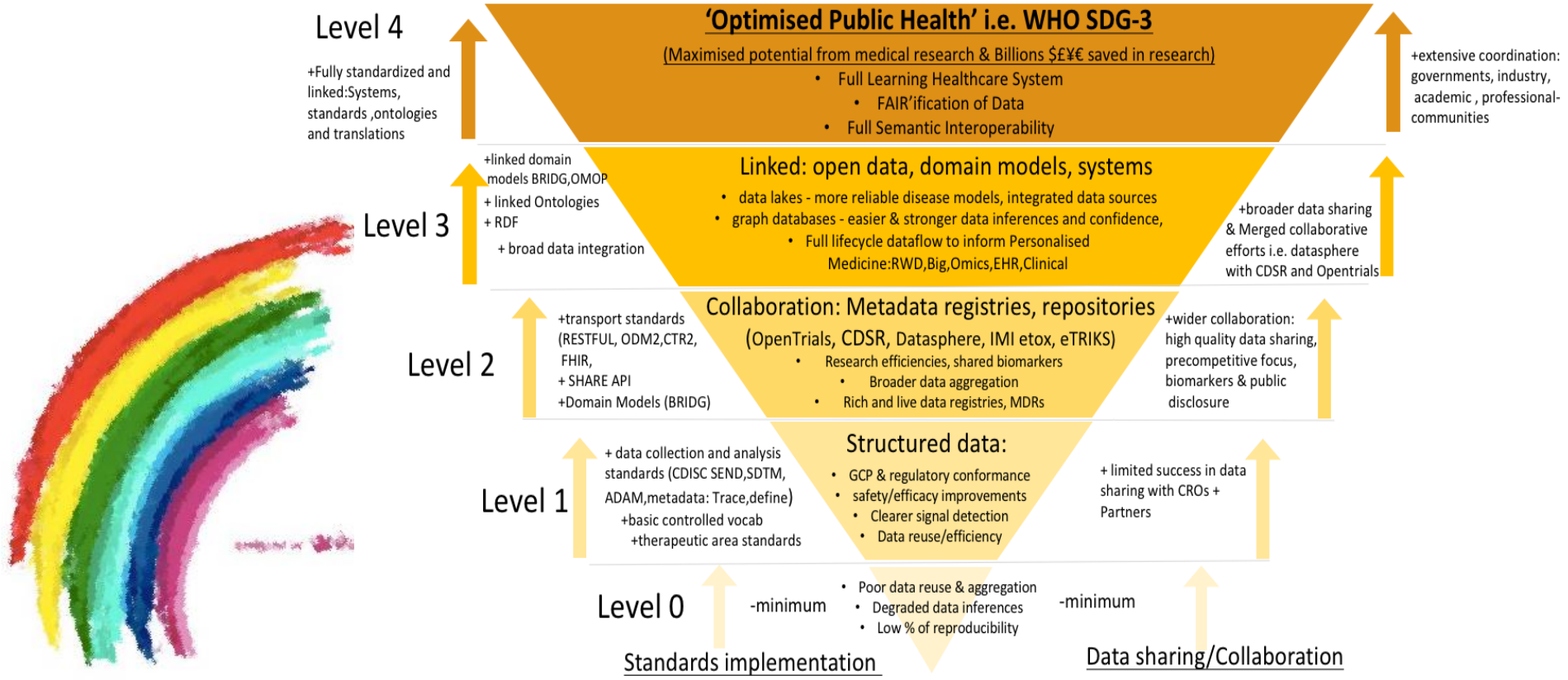
CDISC

Top 5 Recommendations (from the 25 recommendations of the publication)

1. Improve registries via structured and connected data:clinical research & Patient/Disease
2. Merge siloed data sharing efforts through collaborative models.
3. Wider publication and sharing of data concepts and semantic relationships particularly between ontologies ; a formal LOD diagram
4. Increase investment in precompetitive data and knowledge sharing (too many RDF data sets not updated regularly)
5. Validating the representations of the terminology using constraints
   - expressed in RDF Data Shapes using SHACL - extend CDISC RDF

CDISC

Level 4 is an achievable pot of gold at the end of the rainbow



But we all need to be on board with selling the data sharing agenda


Please take our strawman
For data sharing and sell it on…
Apply a lot of heart
…some persistency and dedication
Be courageous and Roar about data Sharing and the benefits

CDISC

With Thanks

- Larry Callahan[2], Michael Braxenthaler[4,] Lauren Becnel[1], Sam Hume[1], Dorina Bratfalean[1], Martin Romaker[3], Philipe Roca-Serra[3,] Andreas Tilman, Susanna Sansone, Ibrahim Emam, Kerstin Forsberg, Frederik Malfait, John Ezzell, Frank Petavy, John Owen, Dave Iberson Hurst
- PhUSE RDF team:
- Phil Ashworth, Scott Bahlavooni, Daniel Boisvert, Susan DeHaven, Nathan Freimark, Josephine-Anne Gough, Laura Hollink, Dave Jordan, Ron Katriel, Kirsten Langendorf, Geoff Low, Frederik Malfait, Mitra Rocca, Gary Walker
- Also recognized are the efforts by Robert Wynne and Erin Muhlbradt

2018
EUROPE INTERCHANGE

CDISC

With Thanks

We would like to thank the CTR2 Team volunteers for their commitment to this project and our sponsors:

ORACLE IPERION Deloitte. HighPoint SOLUTIONS Cognizant accenture